REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	02/24/95	Final 01/01/94 - 12/31/94

4. TITLE AND SUBTITLE
High-Speed Fixed and Floating Point Implementation of Delta-Operator Formulated Discrete Time Systems

5. FUNDING NUMBERS
Grant #:
N00014-94-1-0387
Project Code:
3148509-01

6. AUTHOR(S)
Peter E. Bauer

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dept. of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Code 251 : Jwk
Ballston Tower One
800 N. Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Report was prepared in cooperation with Prof. K. Premaratne, Dept. of Electrical & Computer Engr., Univ. of Miami, Coral Gables, FL 33124

12b. DISTRIBUTION CODE

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

13. ABSTRACT (Maximum 200 words)

This final report describes research results on finite wordlength implementations of delta-operator based discrete systems. It addresses three distinct problems: (a) the existence of limit cycles in fixed and floating point delta-systems, (b) 2-D and m-D delta-system models and (c) extentions of delta-operators to the nonlinear case. All studies in the above three main areas are of comparative nature, i.e. the results are compared with the known results for the shift-operator case and conditions for superiority of the delta-operator are established. In particular, in the first area it was shown, that delta-operator based fixed point designs cannot be free of limit cycles, regardless of the quantization format. It was also shown, that the limit cycle problem is virtually non-existent in floating point realizations, if the mantissa length is sufficiently high. In the second area (b), the 2-D and m-D Roesser model for delta-systems were developed and analyzed. The notions of reachability and observability gramians as well as the notion of balanced realization were introduced and the sensitivity and roundoff noise behavior analyzed. Finally in the third area (c), delta-operator representations of nonlinear systems were developed and analyzed. Sensitivity measure of the state trajectory were developed and evaluated. Quantization error bounds in delta-systems were derived for certain classes of nonlinear functions.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Finite Wordlength, Stability, Quantization Errors, Limit Cycles, Sensitivity, Fixed and Floating Point Arithmetic, Multi-Dimensional Systems , nonlinear systems, chaotic systems, numerical accuracy	252
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Final Report to the

Office of Naval Research (ONR)

for Support of Research Entitled

HIGH SPEED FIXED AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE TIME SYSTEMS

Under the Direction of

Dr. P. H. Bauer, Associate Professor
Department of Electrical Engineering
University of Notre Dame
e-mail: pbauer@mars.ee.nd.edu

Starting date: 01/01/1994
Project duration: 12 months
Date of proposal submission: 09/20/1993
Amount: $39,431.-

Grant Number: N00014-94-1-0387

Contact at ONR: Dr. Clifford G. Lau, Tel. (703) 696-4216
Contact at the University of Notre Dame: Mrs. Ellen Rogers, Tel. (219) 631-7432

Table of Contents

I. Introduction

In many applications such as high speed digital signal processing, reliable simulations of dynamical systems, digital implementation and simulation of chaotic systems, etc., effects of finite wordlength are a critical issue. The process of actual digital computer implementation of a given ideal dynamical system can be characterized by several open parameters that have a critical impact on the performance of the actually implemented algorithm:

1. The realization (in the linear case, the system matrices): This determines the coefficients involved, the order of computation, etc. There are infinitely many realizations for implementing the same dynamical system.
2. The arithmetic format: This determines the type of arithmetic used (fixed point, floating point, etc.), the register lengths, and the type of quantization operations in the reformatting processes.

For the class of linear, time-invariant systems, δ-operator based implementations were shown to perform superior relative to their q-operator based counterparts, if the sampling rate is chosen sufficiently small [1,2]. These advantages of the δ-operator, especially in high speed, real-time applications, were demonstrated with respect to quantization noise at system output and differential sensitivity of the frequency response with respect to coefficients of system realization [1-4]. In addition, the use of δ-operators allows a unified treatment of both the continuous and discrete time cases. These properties make δ-operator based systems an attractive alternative to conventional system realizations.
However, a number of questions on δ-operator based implementations remained unanswered:

1. Do the advantages that have been demonstrated for the linear time-invariant case carry over to the 2-D, m-D, and possibly nonlinear cases? If so, δ-operator based numerical schemes can provide a completely novel, simple, widely applicable, yet more reliable methodology for system simulation and realization. The fundamental importance of such an investigation was identified at the very outset by the PI's.
2. What about asymptotic stability of δ-operator systems and the possibility of limit cycles? Although quantization noise at the output was shown to be smaller for δ-system realizations, this does not automatically preclude the existence of limit cycles. In fixed point implementations, the existence of prohibitively large limit cycles was evident. Although in almost all applications such behavior is unacceptable, no attention had been directed towards this seemingly generic phenomena of δ-systems.

The above questions are at the core of this research project. This report, which provides a description of the work carried out under this research project, is structured as follows: In Section II, a brief description of the proposed tasks is outlined. In Section III, the results obtained are briefly described on a qualitative level for each of the problem areas tackled. Section IV offers conclusions and summarizes the accomplishments and their significance. Section V contains pertinent references. More detailed technical decriptions of the results in Section III may be found in several technical papers, and these are included in Appendix A. It contains all those papers that have already been published in or submitted to journals or conferences as well as all material (such as, presentations, summaries, etc.) that has been submitted to ONR. Appendix B contains those technical papers that have some peripheral relevance to the proposed research, and those in which acknowledgement of ONR support is given.

II. Brief Description Tasks

The proposed work was divided into three major tasks:

T1: Analysis and design of finite wordlength implementations of linear time-invariant δ-systems.
T2: Analysis of nonlinear circuits through δ-operator based schemes.
T3: 2-D and m-D δ-system models.

Task 1 reveals some fundamental difficulties in the implementation of δ-systems with fixed point arithmetic. It focuses mainly on zero convergence of the free system response and exposes the existence of limit cycles as well as effects of sampling time Δ quantization.

Task 2 is a study of whether the superior finite wordlength properties associated with certain linear time-invariant system realizations also extend to nonlinear systems. This work was mainly motivated by some very promising simulation results of chaotic systems.

Task 3 develops the formalisms for 2-D and m-D system descriptions in δ-operator form. It also investigates sensitivity properties of these proposed m-D δ-models and compares them with conventional q-models.

III. Results and Accomplishments

This section offers brief qualitative descriptions of the results obtained during this project period. A more rigorous quantitative analysis of these results are to be found in Appendix A which contains all relevant technical papers.

III.1 Task 1: Analysis and Design of Finite Wordlength Implementations of Linear Time-Invariant δ-Systems

We have exposed a serious limitation of δ-operator based realizations of discrete time systems: they cannot be free of limit-cycles when used with small sampling times and fixed point arithmetic! In particular, DC limit cycles are always present when sampling time is smaller than 0.5 for rounding, and 1.0 for truncation. In other words, under these conditions, nonzero initial conditions can be found, such that the asymptotic response converges to an incorrect equilibrium point different from the origin [5]. This in fact is a generic problem with δ-systems in the sense that it is independent of the margin of stability (of the ideal linear system) and its realization. The main cause of this lies in the update equation where multiplication by sampling time (which is typically small) occurs. This results in a difference vector that quantizes to zero.

The use of novel quantization schemes with smaller deadzones was also shown to be ineffective: Although quantizers that significantly reduce DC limit cycle amplitude may be selected, new oscillatory limit cycles are usually created. A newly developed computer-aided search algorithm for the existence of limit cycles may be effectively used to investigate this phenomenon [6]. Through construction of deadband regions and simple bounding hypercubes, these limit cycle amplitudes have been shown to grow with increasing sampling rate [6,7]. Using results on necessary 1-D conditions for stability of m-D systems [8], m-D δ-systems were also shown to produce similar limit cycle behavior [7,9].

Another drawback of fixed point δ-operator implementations is the required high dynamic range of coefficients and signals. This is due to the fact that, given a q-system, in obtaining the corresponding δ-system, a division by Δ (which is typically small) is involved. Hence, additional bits in coefficient/signal registers are generally required to avoid overflow [10].

The above investigations produced the following unavoidable conclusion: Since δ-operator formulated discrete time systems are superior to their q-operator counterparts only when the sampling rate is chosen to be significantly smaller than one, fixed point arithmetic is not a suitable format for δ-system implementations.

The situation is refreshingly different in floating point arithmetic: The above mentioned problems (encountered under fixed point arithmetic) vanish and δ-systems produce significant advantages under high speed conditions. We show that, under floating point format, a stable linear system (independent of realization) can always be implemented limit cycle free in the regular dynamic range [11]. Equivalently, limit cycles can always be restricted into underflow conditions. Such limit cycles are acceptable for most applications. Furthermore, the large dynamic range requirements of δ-systems may easily be accommodated in floating point arithmetic.

For both fixed and floating point systems, new differential sensitivity measures which are widely applicable even to nonlinear and time-variant systems were developed [10,12]. Instead of using sensitivity measures related to frequency response (as is the usual practice), a time domain approach using state space methods was developed. Sensitivity of state trajectory with respect to system coefficients and initial conditions was investigated. For linear time-invariant systems, δ-operator based implementations have been shown to yield lower (by a factor Δ) sensitivity than their q-operator based counterparts. Sensitivities with respect to initial conditions are shown to be identical for both implementations.

III.2 Task 2: Analysis of Nonlinear Circuits Through δ-Operator Based Schemes

The following aspects of nonlinear δ-systems were addressed in detail:

(a) Sensitivity of state response with respect to coefficients of the nonlinear equation: This analysis was carried out for various types of nonlinearities as well as for both fixed and floating point schemes [10,12].
(b) Bounds on quantization error magnitudes, required dynamic range and construction of majorant systems for the response of δ-operator based implemented nonlinear systems [10].

In part (a), the concept of differential sensitivity of state response with respect to coefficients of the nonlinear equation was developed. The proposed sensitivity measures were evaluated for linear systems, piece-wise linear systems, systems with $\mathcal{C}^1$ nonlinearities and systems with piecewise $\mathcal{C}^1$ nonlinearities.

For all these types of nonlinear systems, sensitivity of a δ-system with respect to coefficients was shown to be smaller (by a factor Δ) than that for its corresponding q-system under fixed point arithmetic. For piece-wise linear and piece-wise $\mathcal{C}^1$ nonlinear systems, development of a quantitative measure for sensitivity of state trajectory with respect to initial conditions was required as well. This is due to how the piecewise characteristics of the nonlinearity is modelled. This proposed sensitivity measure was shown to be comparable for both q- and δ-systems.

Of course, nonlinear δ-systems, implemented in fixed point arithmetic, can be shown to suffer from the same generic problem: Existence of incorrect equilibria. Since this is a serious problem especially in implementation and simulation of nonlinear systems, floating point arithmetic was also intensively analyzed as an alternative. Suitable sensitivity measures were developed and evaluated for the nonlinear system types mentioned above. A comparison with corresponding q-operator based systems revealed what we believe to be a very important observation: Under mild conditions on the coefficients of the q-system, the state trajectory of the corresponding δ-system is less sensitive than that of the q-system. These conditions turn out to be routinely satisfied if the nonlinear discrete time system is obtained through sampling of a given continuous time system with a high sampling rate!

In part (b), a comparison between q- and δ-systems was conducted via quantization error bounds. For the fixed point case, the q-system is always inferior to the δ-system, i.e., it produces larger quantization error bounds whenever single length accumulators are used or when the number of computations in the state equation significantly exceeds the number of computations in the update equation.

Systems with polynomial type nonlinearities has been investigated in great detail. For this class of nonlinearities, recommendations for the sampling rate which would provide an optimal balance between (a) the gains obtained from a reduced sampling period, and (b) the increased expense from a higher sampling frequency, are made. For sector bounded nonlinearities, majorant systems for the state response were constructed. When the sampling time is much smaller than 1, at each time instant, these majorant systems for δ-systems produce smaller state responses than those corresponding to q-systems.

For floating point arithmetic, δ-systems produce smaller quantization error bounds than corresponding q-systems only when the nonlinearities satisfy certain magnitude conditions relative to the state vector. It was shown that, these conditions are always satisfied if the discrete time system is produced by sampling the underlying continuous time system at a very high rate. Note that, this is in accordance with our previous results on sensitivity. The underlying reason for these advantages of δ-systems is due to its implicit operand sorting. In other words, operands of similar 'size' are grouped together in the state equation of δ-systems, whereas, in the q-operator case, such a grouping is not implicit and a mix of operands of different 'sizes' is created.

III.3 Task 3: 2-D and m-D δ-System Models

In this task, the δ-operator counterpart to the 2-D Roesser q-model was developed [13]. It was shown that, for small sampling 'times' in both directions of propagation, the proposed 2-D and m-D models possess similar properties as the 1-D model. For example, fixed point implementations are still plagued by limit cycles and not recommended; however, floating point implementations can yield extremely attractive finite wordlength properties.

The usual system theoretic notions such as characteristic equation, transfer function, stability [14], etc., have been developed for the proposed 2-D δ-models. Furthermore, the notions of gramians, balanced realizations, and also its computation, were introduced.

To investigate coefficient sensitivity properties of the 2-D δ-models, sensitivity measures appropriate for fixed and floating point arithmetic schemes were developed. This analysis was carried out for the more general multi-input, multi-output case. The resulting conclusions may be summarized as follows:

1. In the fixed point case, δ-models yield smaller coefficient sensitivity than the corresponding q-models when the sampling 'times' are small. Balanced realizations exhibit minimum coefficient sensitivity. This parallels the situation encountered in q-operator case. However, note that, generic limit cycle problems persist.

2. In the floating point case, δ-models consistently offer superior coefficient sensitivity when the corresponding q-models' coefficients satisfy certain mild conditions. These conditions are routinely satisfied when the implementing high-Q, digital filters at high speeds. In most situations, 2-4 mantissa bits of an advantage is possible.

Furthermore, computation of balanced realizations has also been addressed. A simple relationship between balanced forms of corresponding q- and δ-system realizations has been established [13]. This makes it possible to derive balanced realizations using those algorithms that are applicable for the q-operator case.

IV. Conclusion

In summary, results obtained during the course of this funding period show that, δ-operator implementations of discrete time systems can be quite superior to their q-operator counterparts if they are used correctly. We have shown that, great gains can be achieved in the case when a continuous time system is sampled at a very high rate and is implemented in floating point arithmetic. Similar comments are applicable to nonlinear and m-D systems as well.

Based on this work, we may make the following conclusion: δ-operator based implementations offer a number of unique and desirable properties which are essential in high performance applications, such as, high speed DSP and reliable simulations of dynamical systems. For such applications (where traditional q-operator based implementations are known to be ill-conditioned), δ-operator based schemes provide a general and easily applicable technique for reliable implementation of discrete time systems.

V. References

[1] G.C. Goodwin, R.H. Middleton and H.V. Poor, "High speed digital signal processing and control," *Proc. IEEE,* Vol. 80, pp. 240-259, 1992.

[2] R. H. Middleton and G. C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Prentice Hall, New Jersey, 1988.

[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proc. IEEE CDC'90,*, Vol. 2 , pp. 954-959, Honolulu, HI, 1990.

[4] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.,* Vol. 41, pp. 629-637, 1993.

[5] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proc. IEEE ISCAS'94,* Vol. 2, pp. 461-464, London, UK, 1994.

[6] K. Premaratne, E.C. Kulasekere, P.H. Bauer, and L.J. Leclerc, "An exhaustive search algorithm for checking limit cycle behavior of digital filters," *IEEE Trans. Sig. Proc.,* in review; a preliminary version: *IEEE ISCAS'95,* Seattle, WA, 1995, to be presented.

[7] P. H. Bauer and K. Premaratne, "Limit cycles in delta-operator formulated 1-D and m-D discrete-time systems with fixed-point arithmetic," *IEEE Trans. Circ. Syst., Pt. I,* in review.

[8] P.H. Bauer, "Low-dimensional conditions for global asymptotic stability of m-D nonlinear digital filters," *Proc. IEEE ISCAS'94,* Vol. 2, pp. 553-556, London, UK, 1994.

[9] P.H. Bauer and K. Premaratne, "Fixed point implementations of m-D delta-operator formulated discrete time systems: Difficulties in convergence," *Proc. IEEE SOUTHEASTCON'94,* Miami, FL, 1994.

[10] P.H. Bauer and K. Premaratne, *Presentation to Dr. C.G. Lau in Sept'94,* ONR, Arlington, VA, 1994.

[11] P.H. Bauer and K. Premaratne, "Zero-convergence of 2-D Roesser state space models implemented in floating-point arithmetic," *38th Midwest Symp. on Circ. and Syst. (MWSCAC'95),* Rio de Janeiro, Brazil, 1995, to be presented.

[12] K. Premaratne and P.H. Bauer, "Digital simulation of nonlinear systems using delta-operator based numerical schemes," *IASTED Int. Conf. Modelling and Simulation,* Colombo, Sri Lanka, 1995, in review.

[13] K. Premaratne, M.M. Ekanayake, J. Suarez, and P.H. Bauer, "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties," *IEEE Trans. Sig. Proc.* in review; preliminary versions: *Proc. 37th Midwest Symp. Circ. Syst. (MWSCAS'94),* Lafayette, LA, 1994; *IEEE SOUTHCON'95,* Fort Lauderdale, FL, 1995, to be presented.

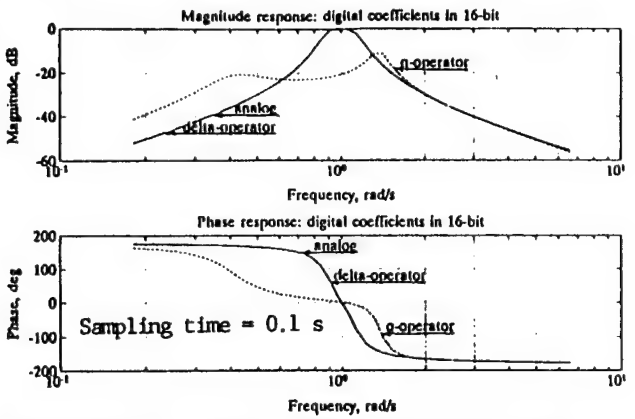
[14] K. Premaratne and A.S. Boujarwah, "An algorithm for stability determination of two-dimensional delta-operator formulated discrete-time systems," *Multidim. Syst. Sig. Proc.,* to appear.

VI. Appendix A: Papers Directly Related to Grant

HIGH-SPEED FIXED- AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS

PETER H. BAUER, University of Notre Dame (Grant No: N00014-94-1-0387)
KAMAL PREMARATNE, University of Miami (Grant No: N00014-94-1-0454)

FIGURE



FREQUENCY RESPONSE OF AN ANALOG PROTOTYPE AND ITS DIGITAL EQUIVALENTS. DIGITAL COEFFICIENTS ARE STORED WITH 16 BITS IN MANTISSA. (Note. Analog and delta-system plots overlap)

APPROACH

*Limit cycle behavior (deadband bounds).
*Coefficient sensitivity (differential sensitivity measures).
*Quantization error (construction of error envelopes).
*Development of balanced forms for 2-D and *m*-D systems.

The approach is applied to three system types:
[T1] Linear, shift-invariant discrete-time systems.
[T2] Digital simulation of nonlinear systems.
[T3] 2-D and *m*-D discrete-time systems.

OBJECTIVE

Application: High performance, real-time applications involving *fast sampling/short wordlength.*

Conventional shift-operator (*q*-operator) based algorithms are ill-conditioned.

Is the delta-operator (δ-operator) based approach more suitable?

*For which classes of systems does it possess better finite wordlength properties?
*Can it improve reliability of computations in simulating nonlinear and multidimensional systems?

ACCOMPLISHMENTS

*With floating-point, δ-systems offer superior performance (especially, for sampled continuous-time systems) with short step size. All fixed-point δ-systems are plagued by limit cycles.
*δ-operator based digital simulation of nonlinear systems offer superior performance (especially, when using a small discretization step size).
*δ-operator models developed for 2-D and *m*-D filters also possess similar properties.

PRELIMINARIES

OPERATORS

- For $\mathbf{x} \in \Re^m$, $q[\cdot]$ is the operator

$$q[\mathbf{x}](n) = \mathbf{x}(n+1).$$

- For $\mathbf{x} \in \Re^m$, $\delta[\cdot]$ is the operator

$$\delta[\mathbf{x}](n) = \frac{\mathbf{x}(n+1) - \mathbf{x}(n)}{\Delta} = \frac{q[\mathbf{x}](n) - \mathbf{x}(n)}{\Delta}.$$

  Here, Δ is a positive constant (usually the sampling time).
- $q[\cdot]$ and $\delta[\cdot]$ are related by

$$q = 1 + \Delta\delta.$$

q-OPERATOR BASED STATE-SPACE MODEL

q-operator based model $\{A_q, B_q, C_q, D_q\}$ of a linear, shift-invariant, causal, p-input, q-output discrete-time system:

$$\begin{aligned} q[\mathbf{x}](n) &= A_q\mathbf{x}(n) + B_q\mathbf{u}(n); \\ \mathbf{y}(n) &= C_q\mathbf{x}(n) + D_q\mathbf{u}(n). \end{aligned}$$

δ-OPERATOR BASED STATE-SPACE MODEL

Corresponding δ-operator based model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$:

Intermediate equation

$$\begin{aligned} \delta[\mathbf{x}](n) &= A_\delta\mathbf{x}(n) + B_\delta\mathbf{u}(n); \\ \mathbf{y}(n) &= C_\delta\mathbf{x}(n) + D_\delta\mathbf{u}(n). \end{aligned}$$

Update equation

$$q[\mathbf{x}](n) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n).$$

- $\{A_q, B_q, C_q, D_q\}$ and $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ are related by

$$A_q = I + \Delta A_\delta; \quad B_q = \Delta B_\delta; \quad C_q = C_\delta; \quad D_q = D_\delta.$$

[T1] LINEAR, SHIFT-INVARIANT DISCRETE-TIME SYSTEMS

*OBJECTIVE

- How do δ-systems perform under fast sampling/short wordlength conditions? What are their properties regarding limit cycles, quantization errors, coefficient sensitivity, and dynamic range?

*ACCOMPLISHMENTS

Both FXP (fixed-point) and FLP (floating-point) schemes are tackled.

*LIMIT CYCLES

- FXP case: δ-systems (with small Δ) always exhibit limit cycles.
- FLP case: Similar to q-systems, with sufficient mantissa length, limit cycles occur only in underflow.

*QUANTIZATION ERROR PROPAGATION

- FXP case: δ-systems possess smaller bounds for quantization after multiplication. Otherwise, both q- and δ-systems are comparable.
- FLP case: In general, δ-systems are better than or equal to q-systems. If δ-system is the digital equivalent of a continuous-time system with fast sampling, it offers superior performance.

*COEFFICIENT SENSITIVITY

- FXP case: δ-systems are superior with fast sampling.
- FLP case: In general, δ-systems are better than or equal to q-systems. If δ-system is the digital equivalent of a continuous-time system with fast sampling, it offers superior performance.

*DYNAMIC RANGE CONSTRAINTS

- FXP case: If δ-system is the digital equivalent of a continuous-time system, both q- and δ-systems are comparable. If a q-system is simply converted to a δ-system, the latter requires a larger dynamic range.
- FLP case: If δ-system is the digital equivalent of a continuous-time system, it is superior. If a q-system is simply converted to a δ-system, the latter requires a slightly larger dynamic range.

LIMIT CYCLES

The ideal linear system is taken to be asymptotically stable. We consider the zero input case.

FXP Case

A δ-system implementation, under finite wordlength, becomes

$$\delta[\mathbf{x}](n) = Q\{A_\delta \mathbf{x}(n)\};$$
$$q[\mathbf{x}](n) = \mathbf{x}(n) + Q\{\Delta \cdot \delta[\mathbf{x}](n)\}.$$

Here, $Q\{\cdot\}$ is the quantization nonlinearity.

Accomplishments

- δ-systems exhibit DC limit cycles if

$$\Delta \leq 0.5 \quad \text{for rounding;} \quad \Delta < 1.0 \quad \text{for truncation.}$$

  Fundamental reason for these limit cycles is the deadzone of quantizer. This creates deadbands for both $\delta[\mathbf{x}]$ and $\mathbf{x}$.
- In fact, limit cycle free δ-system implementations do not exist!
- A smaller sampling time Δ yields a larger deadband for $\delta[\mathbf{x}]$.
- Construction of this deadband for various arithmetic schemes have been performed.
- Structure of system matrix A_δ has a major effect on geometry of deadband for $\mathbf{x}$.
- Reduction of quantizer deadzone reduces size of deadband, thus reducing DC limit cycle amplitude. But, this increases other (oscillatory) limit cycles.
- Neither the use of unconventional quantization nonlinearities nor scaling techniques overcome this difficulty.

FLP Case

Accomplishments

- If mantissa length is sufficiently large, response will always converge into underflow.
- Hence limit cycles may occur only in underflow. This is usually acceptable if dynamic range of underflow is sufficiently small (that is, smallest representable exponent is sufficiently small).

QUANTIZATION ERROR PROPAGATION

Quantization error propagation is investigated via error envelopes.

FXP Case

Accomplishments

- Error envelopes for δ-systems are lower than for corresponding q-systems if quantization occurs after multiplication. Otherwise, they are comparable.

FLP Case

Accomplishments

- In general, error envelopes δ-systems are better than or equal to q-systems.
- However, when q-system matrix A_q is of the form

$$A_q = I + \{\epsilon_{i,j}\},$$

where the matrix elements $\epsilon_{i,j}$ satisfy

$$| \epsilon_{i,j} | << 1,$$

δ-system provides superior performance. This situation occurs, when a digital equivalent of a continuous-time system is obtained with fast sampling.

- In this situation, δ-operator implementation achieves 'operand sorting' (which is known to tremendously reduce quantization errors in FLP realizations).
- Generalized versions of δ-operator, that can tackle situations where A_q does not satisfy the above condition, have been developed. These provide superior performance than q-systems.

COEFFICIENT SENSITIVITY

Coefficient sensitivity is investigated via differential sensitivity measures. Small perturbations are assumed.

- Frequency response sensitivity have been investigated by others.
- Time response or orbit sensitivity arises as a special case of our work in Task [T2] below.

FXP Case

Accomplishments

- δ-systems offer superior performance, in particular, with fast sampling.

FLP Case

Accomplishments

- In general, δ-systems are better than or equal to the corresponding q-systems.
- Conditions under which δ-systems perform better are derived. In particular, if the δ-system is a digital equivalent of a continuous-time system obtained with fast sampling, it offers superior performance.

DYNAMIC RANGE CONSTRAINTS

FXP Case

Accomplishments

- If the δ-system is obtained by discretization of a continuous-time system, the dynamic range requirements of corresponding q- and δ-systems are comparable.
- If the δ-system is obtained by simply converting a q-system, it typically requires a larger dynamic range, larger coefficient registers, and larger accumulators.

FLP Case

Accomplishments

- Wordlength requirements for q- and δ-systems are comparable.
- If the δ-system is obtained by discretization of a continuous-time system with fast sampling, its zero convergence can be guaranteed with less number of bits.

[T2] DIGITAL SIMULATION OF NONLINEAR SYSTEMS

*OBJECTIVE

- Can one perform *reliable* digital simulations of nonlinear systems using δ-operator based numerical schemes?
- If so, just as for linear systems, would one get superior finite wordlength properties?
- The resulting impact and consequences in high performance computing (for example, in digital simulation of nonlinear systems, signal processing, and control) can be significant.

*ACCOMPLISHMENTS

Several important types of nonlinearities were considered.

*LIMIT CYCLES

This is quite similar to the linear case. See our work in Task [T1].

*QUANTIZATION ERROR PROPAGATION

- FXP case: Due to possibility of incorrect equilibria, FXP implementation is not recommended.
- FLP case: Conditions under which δ-systems are superior are derived.

*COEFFICIENT SENSITIVITY

- FXP case: With small grid size, δ-operator based numerical schemes are superior than the conventional q-operator schemes.
- FLP case: Conditions under which coefficient sensitivity of δ-systems are superior are derived. Typical digital equivalents of nonlinear systems under small grid size routinely satisfy these conditions.

*DYNAMIC RANGE CONSTRAINTS

This is quite similar to the linear case. See our work in Task [T1].

q-OPERATOR BASED NONLINEAR SYSTEM

$$q[\mathbf{x}](n) = \mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q).$$

- $\mathbf{a}_q = [a_{1_q}, \ldots, a_{M_q}]^T$ are the coefficients that are *actually stored* in computer.

δ-OPERATOR BASED NONLINEAR SYSTEM

We propose the following:

Intermediate equation

$$\delta[\mathbf{x}](n) = \mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta).$$

Update equation

$$q[\mathbf{x}](n) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n).$$

- $\delta[\mathbf{x}](n) = (q[\mathbf{x}](n) - \mathbf{x}(n))/\Delta$ and $\mathbf{f}_\delta = (\mathbf{f}_q - \mathbf{x})/\Delta$.
- Δ is an arbitrary positive constant (usually the grid size).
- $\mathbf{a}_\delta = [a_{\delta_1}, \ldots, a_{\delta_M}]^T$ are the coefficients that are *actually stored.*

QUANTIZATION ERROR PROPAGATION

FXP Case

Accomplishments

- δ-systems offer superior performance if quantization is performed after multiplication or if polynomial nonlinearities of higher order are to be implemented.
- However, in FXP, δ-systems may converge to incorrect equilibria (see comments in [T1]). Hence, FXP implementation is not recommended.

FLP Case

Accomplishments

- δ-systems show significantly reduced quantization error bounds if $\delta[\mathbf{x}](n) = \mathbf{f}_\delta(\mathbf{x}(n))$ where $\|\Delta \cdot \mathbf{f}_\delta(\mathbf{x}(n))\| << \|\mathbf{x}(n)\|$.
- Under fast sampling, similar to the linear case, this condition is routinely satsified. Hence, δ-operator based discretization schemes, in FLP, can *drastically reduce* quantizations errors with fast sampling.

COEFFICIENT SENSITIVITY

For this presentation, the nonlinearity is taken to belong to $\mathcal{C}^1$, that is, it possesses first partial derivatives. Small perturbations are assumed.

FXP Case

Coefficient perturbation is approximately independent of its nominal value. Hence, a good sensitivity measure of orbit $\mathbf{x}$ is $\partial \mathbf{x}/\partial \mathbf{a}|_n \doteq \partial \mathbf{x}(n)/\partial \mathbf{a}$.

Accomplishments

- Comparison of q- and δ-systems: $\partial \mathbf{x}/\partial \mathbf{a}_q|_n = \Delta \cdot \partial \mathbf{x}/\partial \mathbf{a}_\delta|_n$. Hence, δ-operator based schemes offer superior coefficient sensitivity when Δ is small.
- Similar comments hold true for linear systems, piecewise $\mathcal{C}^1$ nonlinear systems, and piecewise linear systems.

FLP Case

Coefficient perturbation is approximately proportional to its nominal value. Hence, a good sensitivity measure of orbit $\mathbf{x}$ is

$$\frac{\partial \mathbf{x}}{\partial \mathbf{a}/\mathbf{a}}\Big|_n \doteq \begin{bmatrix} \frac{\partial}{\partial a_1/a_1}\mathbf{x}(n) \\ \vdots \\ \frac{\partial}{\partial a_M/a_M}\mathbf{x}(n) \end{bmatrix}.$$

Accomplishments

- Comparison of q- and δ-systems: We have shown that, δ-operator based schemes offer superior coefficient sensitivity if

$$|a_{i_q} - 1| \leq |a_{i_q}|, \ \forall i = 1, \ldots m.$$

  Here, a_{i_q} indicates the 'linear' term in the i-th equation of $\mathbf{f}_q$.
- Similar comments hold true for linear systems, piecewise $\mathcal{C}^1$ nonlinear systems, and piecewise linear systems.

Example: Lorenz Equation

Consider the digital simulation of Lorenz equation:

$$x_1^{(1)}(t) = a_{11}x_1(t) + a_{12}x_2(t);$$
$$x_2^{(1)}(t) = a_{21}x_1(t) + a_{22}x_2(t) + a_{213}x_1(t)x_3(t);$$
$$x_3^{(1)}(t) = a_{33}x_3(t) + a_{312}x_1(t)x_2(t).$$

Here, $a_{11} = -\sigma$, $a_{12} = \sigma$, $a_{21} = \rho$, $a_{22} = -1$, $a_{213} = -1$, $a_{33} = -\beta$, and $a_{312} = 1$.

q-operator based forward Euler scheme with $\Delta = 1e-04$

$$q[x_{1_q}](n) = a_{11_q}x_{1_q}(n) + a_{12_q}x_{2_q}(n);$$
$$q[x_{2_q}](n) = a_{21_q}x_{1_q}(n) + a_{22_q}x_{2_q}(n) + a_{213_q}x_{1_q}(n)x_{3_q}(n);$$
$$q[x_{3_q}](n) = a_{33_q}x_{3_q}(n) + a_{312_q}x_{1_q}x_{2_q}(n).$$

Here, $a_{11_q} = 1-\Delta\sigma$, $a_{12_q} = \Delta\sigma$, $a_{21_q} = \Delta\rho$, $a_{22_q} = 1-\Delta$, $a_{213_q} = -\Delta$, $a_{33_q} = 1 - \Delta\beta$, and $a_{312_q} = \Delta$.

δ-operator based forward Euler scheme with $\Delta = 1e-04$

$$\delta[x_{1_\delta}](n) = a_{11_\delta}x_{1_\delta}(n) + a_{12_\delta}x_{2_\delta}(n);$$
$$\delta[x_{2_\delta}](n) = a_{21_\delta}x_{1_\delta}(n) + a_{22_\delta}x_{2_\delta}(n) + a_{213_\delta}x_{1_\delta}(n)x_{3_\delta}(n);$$
$$\delta[x_{3_q}](n) = a_{33_\delta}x_{3_\delta}(n) + a_{312_\delta}x_{1_\delta}x_{2_\delta}(n).$$

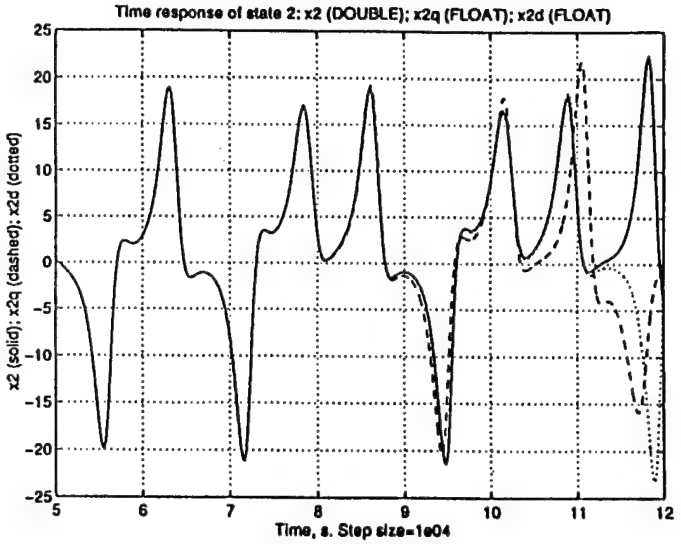
Here, $a_{11_\delta} = a_{11}$, $a_{12_\delta} = a_{12}$, $a_{21_\delta} = a_{21}$, $a_{22_\delta} = a_{22}$, $a_{213_\delta} = a_{213}$, $a_{33_\delta} = a_{33}$, and $a_{312_\delta} = a_{312}$.

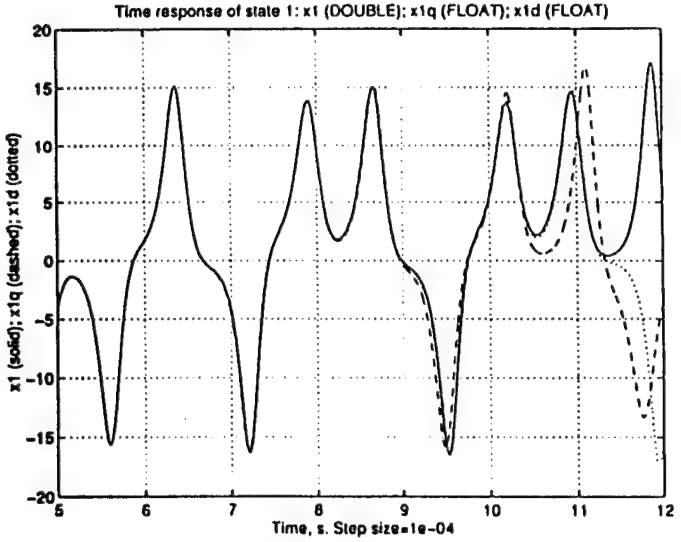
Simulation data

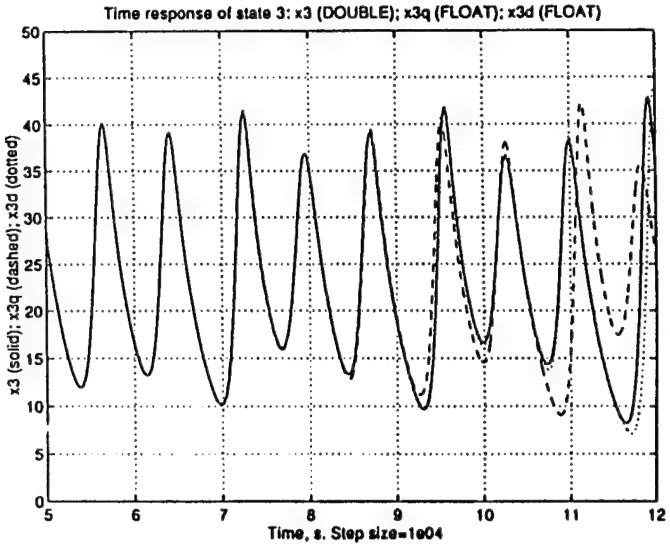
- Nominal coefficient values: $\sigma = 10$; $\rho = 28$; $\beta = 8/3$. This system exhibits chaotic behavior.
- Initial conditions: $\mathbf{x}_q(0) = \mathbf{x}_\delta(0) = [0, 5, 75]^T$.
- Data type: Two simulations were implemented in C using both FLOAT (32-bit FLP) and DOUBLE (64-bit FLP) data types.
- Comparison: DOUBLE simulations until 8 s (where both q- and δ- DOUBLE schemes are identical) were taken as a benchmark for comparison of FLOAT simulations. Clearly, the computed orbit from the δ-scheme is more reliable for a longer duration!

State responses of the Lorenz equation using q- and δ-operator based integration schemes with DOUBLE (64-bit FLP) and FLOAT (32-bit FLP) data types.

1. **Nominal coefficient values:** $\sigma = 10$; $\rho = 28$; $\beta = 8/3$.
2. **Initial conditions:** $\mathbf{x}_q(0) = \mathbf{x}_\delta(0) = [0, 5, 75]^T$.
3. **Integration scheme:** Forward Euler with step size $\Delta = 1e-04$.
4. **Coefficients being stored:**
   - 4.1. q-operator based scheme: $\mathbf{x}_q(0)$; a_{11_q}; a_{12_q}; a_{21_q}; a_{22_q}; a_{213_q}; a_{33_q}; a_{312_q}.
   - 4.2. δ-operator based scheme: $\mathbf{x}_\delta(0)$; a_{11_δ}; a_{12_δ}; a_{21_δ}; a_{22_δ}; a_{213_δ}; a_{33_δ}; a_{312_δ}; Δ.
5. **Data type:**
   - 5.1. DOUBLE q- and δ-schemes: Both schemes are identical until approximately 28 s. These are shown as 'solid' lines.
   - 5.2. FLOAT q-scheme: These are shown as 'dashed' lines.
   - 5.3. FLOAT δ-scheme: These are shown as 'dotted' lines.







[T3] 2-D AND m-D DISCRETE-TIME SYSTEMS

*OBJECTIVE

- Do the superior finite wordlength properties hold true if 2-D and m-D discrete-time systems are implemented using δ-operator?
- If so, such implementations are useful in high performance, real-time applications that use fast sampling/short wordlength.

*ACCOMPLISHMENTS

*FUNDAMENTAL SYSTEM THEORETIC CONCEPTS

- δ-operator analog of the 2-D Roesser q-model.
- Notions of characteristic equation, transfer function, stability, etc., have been developed.
- Algorithm to check stability, notions of gramians and balanced realizations have been developed.

*COEFFICIENT SENSITIVITY

- FXP case: Balanced realizations possess 'minimum' coefficient sensitivity.
- FLP case: Conditions under which δ-systems perform better are derived. Typically, narrowband high speed digital filters satisfy these requirements.

FUNDAMENTAL SYSTEM THEORETIC CONCEPTS

Operators

- Define operators $q_h[\cdot]$ and $q_v[\cdot]$ as

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1).$$

- Propose operators $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as

$$\delta_h[\mathbf{x}](i,j) = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h};$$
$$\delta_v[\mathbf{x}](i,j) = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v}.$$

Here, Δ_h and Δ_v are positive constants (that are the counterparts of sampling time).

q-Oerator Based Roesser Model

q-operator based Roesser model $\{A_q, B_q, C_q, D_q\}$ of a linear, shift-invariant, strictly causal, p-input, q-output 2-D discrete-time system:

$$\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} = \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j)$$

$$\doteq [A_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_q]\mathbf{u}(i,j);$$

$$\mathbf{y}(i,j) = \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j)$$

$$\doteq [C_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j).$$

δ-Operator Based Roesser Model

We propose the following δ-operator based Roesser model:

Intermediate equation

$$\begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} = \begin{bmatrix} A_\delta^{(1)} & A_\delta^{(2)} \\ A_\delta^{(3)} & A_\delta^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_\delta^{(1)} \\ B_\delta^{(2)} \end{bmatrix} \mathbf{u}(i,j)$$

$$\doteq [A_\delta] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_\delta]\mathbf{u}(i,j);$$

$$\mathbf{y}(i,j) = \begin{bmatrix} C_\delta^{(1)} & C_\delta^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_\delta]\mathbf{u}(i,j)$$

$$\doteq [C_\delta] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_\delta]\mathbf{u}(i,j).$$

Update equation

$$q_h[\mathbf{x}^h](i,j) = \mathbf{x}^h(i,j) + \Delta_h \cdot \delta_h[\mathbf{x}^h](i,j);$$
$$q_v[\mathbf{x}^v](i,j) = \mathbf{x}^v(i,j) + \Delta_v \cdot \delta_v[\mathbf{x}^v](i,j).$$

- $\{A_q, B_q, C_q, D_q\}$ and $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ are related by

$$A_q = I + \tau A_\delta; \quad B_q = \tau B_\delta; \quad C_q = C_\delta; \quad D_q = D_\delta.$$

Here, $\tau = [\Delta_h I \oplus \Delta_v I]$.

Gramians

Analogous to the 1-D and 2-D q-operator cases, reachability and observability gramian are proposed as:

$$P \doteq \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} FF^* \frac{dc_h}{1+\Delta_h c_h} \frac{dc_v}{1+\Delta_v c_v};$$

$$Q \doteq \begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} G^*G \frac{dc_h}{1+\Delta_h c_h} \frac{dc_v}{1+\Delta_v c_v}.$$

Here, $F(c_h, c_v) \doteq (I - A_\delta)^{-1} B_\delta$ and $G(c_h, c_v) \doteq C_\delta (I - A_\delta)^{-1}$. $\mathcal{T}_\delta^2$ denotes stability boundary.

Balanced Realizations

It is proposed to call $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ *balanced* if

$$P^{(1)} = Q^{(1)} = \text{diag}\{\sigma_1^{(1)}, \ldots, \sigma_{n_h}^{(1)}\};$$
$$P^{(4)} = Q^{(4)} = \text{diag}\{\sigma_1^{(4)}, \ldots, \sigma_{n_v}^{(4)}\}.$$

Accomplishments

- Characteristic equation and transfer function, relationship with q-model, equivalent transformations, algorithm to check stability, etc., are developed.
- Computation of gramians is addressed. For separable systems, they are block diagonal and may be computed via solution of four Lyapunov equations.

COEFFICIENT SENSITIVITY

Coefficient sensitivity of proposed model is investigated via suitable differential sensitivity measures. Small perturbations are assumed.

FXP Case

Coefficient perturbation is approximately independent of its nominal value. Hence, define

$$M_{\mathrm{FXP}} \doteq \|S_{A_\delta}\|_1^2 + \frac{1}{p}\|S_{B_\delta}\|_2^2 + \frac{1}{q}\|S_{C_\delta}\|_2^2 + \frac{1}{pq}\|S_{D_\delta}\|_2^2.$$

Here, $S_{A_\delta} = \partial H_\delta / \partial A_\delta$, etc. H_δ is the transfer function.

Accomplishments

- Realizations that are bound optimal with respect to M_{FXP} are in fact balanced.
- When $\Delta_h < 1$ and $\Delta_v < 1$, that is, with fast 'sampling', balanced δ-model is better than its corresponding q-model.

FLP Case

Coefficient perturbation is approximately proportional to its nominal value. Hence, define

$$M_{\mathrm{FLP}} \doteq \|\tilde{S}_{A_\delta}\|_1^2 + \frac{1}{p}\|\tilde{S}_{B_\delta}\|_2^2 + \frac{1}{q}\|\tilde{S}_{C_\delta}\|_2^2 + \frac{1}{pq}\|\tilde{S}_{D_\delta}\|_2^2.$$

Here, $\tilde{S}_{A_\delta} = \sum\sum a_{ij_\delta} \partial H_\delta / \partial a_{ij_\delta}$, etc.

Accomplishments

- Realization that are bound optimal with respect to M_{FLP} are better than its corresponding q-model if

$$\|A_q - I\|_F^2 < \|A_q\|_F^2.$$

- High speed narrowband digital filters typically satisfy this requirement.

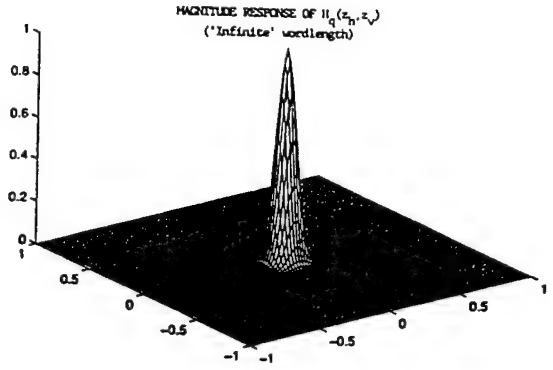
Example: Narrowband 5h-5v 2-D separable digital filter

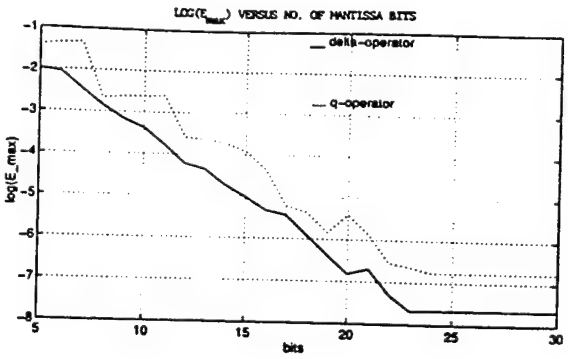
The corresponding q-Roesser model and transfer function are denoted as $\{A_q, B_q, C_q, D_q\}$ and $H_q(z_h, z_v)$, respectively.

- Let $\{\tilde{A}_q, \tilde{B}_q, \tilde{C}_q, \tilde{D}_q\}$ denote the corresponding balanced q-system. Under finite wordlength, let the transfer function be $\tilde{H}_q(z_h, z_v)$.
- Let $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ denote the corresponding balanced δ-system. Under finite wordlength, let the transfer function be $\tilde{H}_\delta(c_h, c_v)$.

Simulation Data

- Mantissa length: $\tilde{H}_q$ and $\tilde{H}_\delta$ were implemented with different mantissa lengths (for the coefficients) to see the effects of coefficient sensitivity.
- Plot shows $\log[E_{\max}]$ versus mantissa length. Here, $\log[E_{\max}] = |\tilde{H}_q - H_q|$ (for the q-system) or $\log[E_{\max}] = |\tilde{H}_\delta - H_q|$ (for the δ-system). H_q is implemented with 'infinite' wordlength.
- Clearly, balanced δ-system performs better than the balanced q-system (which is 'optimal' with respect to the q-system counterpart of sensitivity measure M_{FXP})!





COMPARISON OF IMPLEMENTATIONS

	FXP CASE	FLP CASE
Quantization error bounds General system	δ-systems mostly superior	δ-systems better than or equal to q-systems
Quantization error bounds Digital equivalent of continuous-time system with short sampling time	δ-systems mostly superior	δ-systems superior
Limit cycles	δ-systems exhibit limit cycles	q- and δ-systems both exhibit limit cycles only in underflow
Dynamic range constraints Register overflow	δ-systems more likely to cause overflow	Unlikely in both q- and δ-systems
Coefficient sensitivity General system	δ-systems superior	δ-systems better than or equal to q-systems
Coefficient sensitivity Digital equivalent of continuous-time system with short sampling time	δ-systems superior	δ-systems superior
Hardware requirements Implementation of δ^{-1} requires additional sum and product	δ-systems require longer registers (for both coefficients and signals)	q- and δ-systems comparable

PROJECT TITLE:

High-speed fixed- and floating-point implementation of delta-operator formulated discrete-time systems

PRINCIPAL INVESTIGATORS:

• **Kamal Premaratne**

University of Miami.

Grant No: N00014-94-1-0454; R&T Project Code: 3148508—01.

• **Peter H. Bauer**

University of Notre Dame.

Grant No: N00014-94-1-0387; R&T Project Code: 3148509—01.

SUMMARY OF PHASE P2 RESULTS

The work described in this report is related to the following

[T2] Task T2: Analysis of nonlinear circuits through δ-operator based schemes.

Problems Posed in Task T2:

Regarding the proposals associated with the above grants, within Task T2, the following questions were raised:

1. With the desirable properties of δ-systems applicable to linear systems in mind, does the same carry over if nonlinear systems are implemented with δ-operator based schemes?

2. In particular, issues concerning coefficient sensitivity and quantization noise is of special importance in such systems.

3. If a δ-operator based scheme offers significant improvements over its q-operator counterpart, the consequences in nonlinear signal processing, nonlinear control, and digital simulation of nonlinear dynamics can be significant. .

In fact, the superior finite wordlength performance of the discrete simulation of Chua's Circuit in the grant proposals using the δ-operator, instead of the more conventional q-operator, provided the impetus for the work proposed in Task T2. The work described herein justifies our preliminary optimism and show that this superior performance can be expected with δ-operator based implementations.

This task was proposed to be carried out during Phase P2 with close collaboration between the two PI's. During the whole project duration, both PI's have been in constant contact. In particular, a considerable portion of the work described herein was seen to maturity during a one-week research stay at University of Notre Dame during August 09-16, 1994. During this time, important results that address coefficient sensitivity and quantization error bounds applicable to δ-operator based implementation of nonlinear systems were developed. A description of those Phase P2 results pertaining to coefficient sensitivity follows.

Task T2: Results Pertaining to Coefficient Sensitivity—Summary

Briefly, conclusions drawn from this work may be summarized as follows: We have investigated orbits of linear and nonlinear systems. Several important types of nonlinearities—$\mathcal{C}^1$ nonlinearities, piecewise $\mathcal{C}^1$ nonlinearities, and piecewise linear—were looked into.

- The Fixed-Point Arithmetic (FXP) Case:

With small step size, δ-systems provide superior coefficient sensitivity performance.

- The Floating-Point Arithmetic (FLP) Case:

Conditions under which δ-systems provide superior coefficient sensitivity were derived. Typical digital equivalents of nonlinear systems derived for simulation purposes in fact routinely satisfy these conditions when the step size is small.

Task T2: Results Pertaining to Coefficient Sensitivity—Brief Description

Consider the following q-operator based implementation of a nonlinear system:

$$q[\mathbf{x}](n) = \mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q), \tag{1}$$

where $q[\mathbf{x}] = \mathbf{x}(n+1)$. Here, $\mathbf{x}(n)$ is the state $\mathbf{x} \in \Re^m$ at time instant n and $\mathbf{a}_q = [a_{q_1}, \ldots, a_{q_M}]^T \in \Re^M$ refers to the system parameters that are *actually stored* within the computer.

The corresponding δ-operator based scheme of the same nonlinear system is of the form

$$\begin{aligned} \delta[\mathbf{x}](n) &= \mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) \quad \text{(Intermediate equation)} \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n) \quad \text{(Update equation)} \end{aligned} \tag{2}$$

where $\delta[\mathbf{x}](n) = (q[\mathbf{x}](n) - \mathbf{x}(n))/\Delta$ and

$$\mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) = \frac{\mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q) - \mathbf{x}(n)}{\Delta}. \tag{3}$$

Here, $\Delta \in \Re$ is an arbitrary positive real parameter and $\mathbf{a}_\delta = [a_{\delta_1}, \ldots, a_{\delta_M}]^T \in \Re^M$ again refers to the system parameters that are *actually stored* within the computer.

To see the relationship between $\mathbf{a}_q$ and $\mathbf{a}_\delta$, let the i-th equation in (1) be

$$q[x_i](n) = f_{q_i}(x_1(n), \ldots, x_m(n), a_{q_1}, \ldots, a_{q_M}),\ i = 1, \ldots, m. \tag{4}$$

Then, we may encounter one of two situations:

1. There is a linear term corresponding to x_i, that is, a term of the nature $a_K x_i(n)$, on the RHS of (4). Then, we need to store

$$a_{\delta_i} = \begin{cases} \frac{a_{q_i}}{\Delta}, & \text{for } i = \{1, \ldots, M\} \setminus K; \\ \frac{a_{q_K} - 1}{\Delta}, & \text{for } i = K. \end{cases} \tag{5}$$

2. There is no linear term corresponding to x_i on the RHS of (4). Then, we need to store

$$a_{\delta_i} = \frac{a_{q_i}}{\Delta}, \text{ for } i = \{1, \ldots, M\}, \quad \text{and} \quad \frac{1}{\Delta}. \tag{6}$$

Remark.

1. Of course, in an infinite wordlength implementation, there simply is no difference

between the q- and δ-operator based schemes in (1) and (2). In fact, the latter requires a modest increase in the number of computations. However, what we address is the performance under finite wordlength high-speed conditions.

2. Discretization of a nonlinear system of the form

$$\mathbf{x}^{(1)}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{a}) \tag{7}$$

can give rise to equations of the type in (1) and (2). Here, $\mathbf{x}^{(i)}$ is the i-th derivative of $\mathbf{x}$.

3. In what follows, $\mathbf{f}(\mathbf{x}, \mathbf{a}) \in \mathcal{C}^1$ denotes a nonlinear function that possesses first partial derivatives.

Now, which of the schemes (1) or (2) yield superior coefficient sensitivity properties of its orbit with respect to perturbations of $\mathbf{a}_q$ or $\mathbf{a}_\delta$, respectively? This consideration is crucial in high-speed applications where a shorter wordlength is the avenue of choice.

In what follows, the following standing assumptions are made:

1. All perturbations are small.

2. Comparison between q- and δ-operator based implementations are done with respect to upper bounds (constructed through appropriate norms) on possible errors due to coefficient sensitivity.

FXP CASE

In the FXP case, a good indication of the coefficient sensitivity of the orbit $\mathbf{x}$ is its first partial derivative with respect to the stored coefficient vector $\mathbf{a}$, that is,

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}}\right|_n \doteq \frac{\partial}{\partial \mathbf{a}}\mathbf{x}(n) \in \Re^{mM}. \tag{8}$$

I. $\mathcal{C}^1$ nonlinear system

q-operator case

For this case, we can show the following

THEOREM 1. For the q-operator based implementation in (1),

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_q}\right|_{n+1} = \sum_{j=0}^{n-1}\left(I_M \otimes \prod_{i=j+1}^{n}\left[\begin{array}{ccc}\frac{\partial \mathbf{f}_q}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}_q}{\partial x_m}\end{array}\right]\Big|_i\right)\cdot\left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q}\right|_j + \left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q}\right|_n,$$

where I_M denotes the identity matrix of size $M \times M$ and

$$\left[\begin{array}{ccc}\frac{\partial \mathbf{f}_q}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}_q}{\partial x_m}\end{array}\right]\Big|_i \doteq \left[\begin{array}{ccc}\frac{\partial}{\partial x_1}\mathbf{f}_q(i) & \cdots & \frac{\partial}{\partial x_m}\mathbf{f}_q(i)\end{array}\right] \in \Re^{m\times m};$$

$$\left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q}\right|_j \doteq \frac{\partial}{\partial \mathbf{a}_q}\mathbf{f}_q(j) \in \Re^{mM}.$$

δ-operator case

For brevity, we only consider the case in (5). Note that,

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_\delta}\right|_n = \left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_q}\right|_n \cdot \Delta. \tag{9}$$

In addition, we need to consider the sensitivity of the orbit with respect to Δ (due to the update equation in (2)). However, if we assume an exact FXP representation for Δ, this term could be ignored.

Using, for instance, a norm to compare the sensitivity measures, we conclude that, the δ-operator based implementation will provide superior coefficient sensitivity performance.

Remark. We obtain similar results when considering the case in (6). Here, one may need to consider sensitivity with respect to $1/\Delta$ as well. Again, we may assume that $1/\Delta$ (and Δ) have exact FXP representations. Even if this is not the case, δ-system is still likely to be superior since the reduction in sensitivity gained through other terms is Δ-fold.

II. Linear system

The superior coefficient sensitivity of the frequency response of δ-operator based systems is thoroughly investigated in Li and Gevers (1990). However, no result exists that address the coefficient sensitivity of the *orbit.*

q-operator case

With the more general result in Theorem 1, we can show the following

THEOREM 2. For the q-operator based implementation $\mathbf{x}(n+1) = A_q\mathbf{x}(n)$, $A_q \in \Re^{m\times m}$,

$$\frac{\partial \mathbf{x}}{\partial A_q}\Big|_{n+1} = \sum_{j=0}^{n-1}(I_m \otimes A_q^{n-j})\overline{U}_{m\times m}(I_m \otimes \mathbf{x})\Big|_j + \overline{U}_{m\times m}(I_m \otimes \mathbf{x})\Big|_n,$$

where $\overline{U}_{q\times p} = \sum_{i=1}^{q}\sum_{j=1}^{p} E_{ij}^{(q\times p)} \otimes E_{ij}^{(q\times p)} \in \Re^{q^2\times p^2}$, $E_{ij}^{(q\times p)} = \mathbf{e}_i^{(q)}\mathbf{e}_j^{(p)^T} \in \Re^{q\times p}$. Here, $\mathbf{e}_i^{(n)} \in \Re^n$ is the unit vector with 1 on its i-th row (Brewer 1978).

δ-operator case

The corresponding δ-system's intermediate equation is $\delta[\mathbf{x}](n) = A_\delta\mathbf{x}(n)$ where $A_\delta = (A_q - I)/\Delta$. The update equation of course is as in (2).

Again, as in Section I, we can show that, the δ-operator based implementation will provide superior coefficient sensitivity performance.

III. Piecewise $\mathcal{C}^1$ nonlinear system

Consider a nonlinearity that is piecewise and possesses first partial derivatives within each 'piece'. To address its coefficient sensitivity, we model the dynamics of such a system as follows:

1. Within each 'piece', the system dynamics is a $\mathcal{C}^1$ nonlinearity.

2. Each instant of the orbit's 'entry' into another 'piece' is modeled as a perturbation in the initial conditions.

Regarding item 1, as previous results indicate, the δ-operator based implementation will be superior within each 'piece'. Regarding item 2, we need to investigate the orbit's coefficient sensitivity with respect to initial conditions. This is addressed now.

q-operator case

A reasonable sensitivity measure is

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}(0)}\Big|_n \doteq \frac{\partial}{\partial \mathbf{x}(0)}\mathbf{x}(n) \in \Re^{m^2}. \tag{10}$$

Then, we can show the following

THEOREM 3. For the q-operator based implementation in (1),

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{x}(0)}\right|_{n+1} = I_m \otimes \prod_{i=0}^{n} \left[\frac{\partial \mathbf{f}_q}{\partial x_1} \quad \cdots \quad \frac{\partial \mathbf{f}_q}{\partial x_m} \right]\Big|_i \cdot \quad \frac{\partial \underline{x}(0)}{\partial \underline{x}(0)}$$

δ-operator case

One may show that, Theorem 3 is equally applicable for the δ-operator case as well.

Hence, regarding sensitivity due to initial conditions, both q- and δ-operator based implementations are expected to be provide comparable results.

This implies that, in totality, δ-operator based implementations will provide superior results.

IV. Piecewise linear system

Again, we address the coefficient sensitivity of the orbit with respect to the initial conditions.

q-operator case

As in Section II, with the more general result in Theorem 3, we can show the following

THEOREM 4. For the q-operator based implementation of $\mathbf{x}(n+1) = A_q \mathbf{x}(n)$,

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{x}(0)}\right|_{n+1} = A_q^{n+1}. \quad \text{vec} \quad \Lambda$$

δ-operator case

Again, one may show that, Theorem 4 is equally applicable for the δ-operator case as well.

Hence, as in Section III, δ-operator based implementations will provide superior results.

FLP CASE

In the FLP case, representable values are spaced farther apart at higher values of the parameter. Hence, instead of that used for the FXP case (see (8)), a more realistic sensitivity measure is (see Li and Gevers (1990))

$$\frac{\partial \mathbf{x}}{\partial \mathbf{a}/\mathbf{a}}\Big|_n \doteq \begin{bmatrix} \frac{\partial}{\partial a_1/a_1}\mathbf{x}(n) \\ \vdots \\ \frac{\partial}{\partial a_M/a_M}\mathbf{x}(n) \end{bmatrix} \in \Re^{mM}. \tag{11}$$

I. $\mathcal{C}^1$ nonlinear system

q-operator case

For this case, we can show the following

THEOREM 5. For the q-operator based implementation in (1),

$$\frac{\partial \mathbf{x}}{\partial \mathbf{a}_q/\mathbf{a}_q}\Big|_{n+1} = \sum_{j=0}^{n-1}\left(I_M \otimes \prod_{i=j+1}^{n}\left[\frac{\partial \mathbf{f}_q}{\partial x_1} \quad \cdots \quad \frac{\partial \mathbf{f}_q}{\partial x_m}\right]\Big|_i\right)\cdot \frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q/\mathbf{a}_q}\Big|_j + \frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q/\mathbf{a}_q}\Big|_n.$$

δ-operator case

Again, we only consider the case in (5). Also, let us assume that, the elements in $\mathbf{a}_q$ are enumerated such that, for each $i = 1, \ldots, m$, a_i is the 'linear' element of the i-th equation. Then, note that,

$$\begin{aligned} \frac{\partial \mathbf{x}}{\partial \mathbf{a}_\delta/\mathbf{a}_\delta}\Big|_n &= \begin{bmatrix} a_{\delta_1}\frac{\partial \mathbf{x}}{\partial a_{\delta_1}} \\ \vdots \\ a_{\delta_M}\frac{\partial \mathbf{x}}{\partial a_{\delta_M}} \end{bmatrix}\Big|_n \\ &= \begin{bmatrix} (a_{q_1}-1)\frac{\partial \mathbf{x}}{\partial a_{q_1}} \\ \vdots \\ (a_{q_m}-1)\frac{\partial \mathbf{x}}{\partial a_{q_m}} \\ a_{q_{m+1}}\frac{\partial \mathbf{x}}{\partial a_{q_{m+1}}} \\ \vdots \\ a_{q_M}\frac{\partial \mathbf{x}}{\partial a_{q_M}} \end{bmatrix}\Big|_n, \end{aligned} \tag{12}$$

where we have used (5) and (9). As before, we may ignore the effect of Δ.

Again, we use a norm to compare the sensitivity measures. For instance, using the 1- or ∞-norm, we conclude that, the δ-operator based implementation will provide superior

coefficient sensitivity performance if

$$|a_{q_i} - 1| \leq |a_{q_1}|\,, \ \forall i = 1, \ldots, m. \tag{13}$$

But, how practical is this restriction? In other words, how often, if at all, is it satisfied in practice? To address this, consider the following

Example. Lorenz equation. Consider the state-space description of the Lorenz equation:

$$\begin{aligned} x_1^{(1)}(t) &= -\sigma\,(x_1(t) - x_2(t)); \\ x_2^{(1)}(t) &= \rho\, x_1(t) - x_2(t) - x_1(t)x_3(t); \\ x_3^{(1)}(t) &= x_1(t)x_2(t) - \beta\, x_3(t). \end{aligned}$$

For digital simulation of the corresponding orbit, we use the forward Euler scheme with step size Δ. This yields

$$\begin{aligned} x_1(n+1) &= (1 - \Delta\sigma)\, x_1(n) + \Delta\sigma\, x_2(n); \\ x_2(n+1) &= \Delta\rho\, x_1(n) + (1 - \Delta)\, x_2(n) - \Delta\, x_1(n)x_3(n); \\ x_3(n+1) &= (1 - \Delta\beta)\, x_3(n) + \Delta\, x_1(n)x_2(n). \end{aligned}$$

We at once observe the following: For a small step size Δ,

1. Linear terms are close to 1.

2. Other terms are very small.

Hence, the condition in (13) is in fact satisfied!

In fact, when digital simulation of nonlinear systems are carried out, (13) is often satisfied for a small step size (which denotes fast sampling). Hence, we conclude that, a δ-operator based implementation of such a simulation will provide superior coefficient sensitivity performance!!

II. Linear system

Again, no result that addresses coefficient sensitivity of the *orbit* of linear systems implemented using FLP arithmetic is available.

Without delving into much detail, we simply state the relevant result: Consider the q-operator based implementation $\mathbf{x}(n+1) = A_q\mathbf{x}(n)$ and its corresponding δ-operator based implementation. With respect to the FLP coefficient sensitivity measure introduced

above, the coefficient sensitivity of the δ-system is superior (in terms of the norm being used) than that for the corresponding q-system if

$$\|A_q - I\| \leq \|A_q\|. \tag{14}$$

It is not hard to show the following:

$$|\lambda_i[A_q] - 1| \leq |\lambda_i[A_q]|, \ \forall i = 1, \ldots, m \iff \|A_q - I\|_F \leq \|A_q\|_F; \tag{15a}$$

$$|\lambda_i[A_q] - 1| \leq |\lambda_j[A_q]|, \ \forall i, j = 1, \ldots, m \iff \|A_q - I\|_2 \leq \|A_q\|_2. \tag{15b}$$

$$|\mathrm{diag}_i[A_q] - 1| \leq |\mathrm{diag}_i[A_q]|, \ \forall i = 1, \ldots, m \iff \|A_q - I\|_{1,\infty} \leq \|A_q\|_{1,\infty}. \tag{15c}$$

Here, $\lambda_i[A_q]$ denotes the i-th eigenvalue of A_q and $\mathrm{diag}_i[A_q]$ denotes the i-th diagonal element of A_q.

Hence, if any one of the above conditions are satisfied, the δ-operator based implementation will provide superior coefficient sensitivity performance.

Remark. Li and Gevers (1993) refers to the region in condition (15a) as the *Middleton-Goodwin (MG) Region.* They have shown that, if (15a) is satisfied, the δ-operator based implementation will provide superior coefficient sensitivity of its *frequency response.*

Regarding systems corresponding to those in Sections III and IV of the FXP case, δ-systems offer similar advantages.

Example, continued. Lorenz equation. To justify and validate the results above, a digital simulation of the Lorenz equation was carried out using both q- and δ-operator schemes with FLP. The results are summarized in the series of graphs.

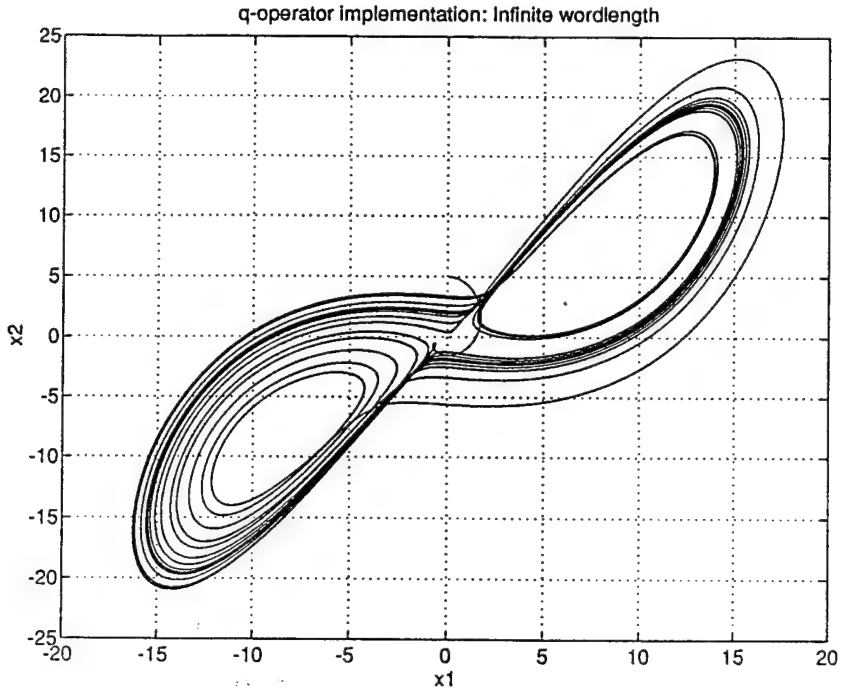
1. Nominal coefficient values: $\sigma = 10$; $\rho = 28$; $\beta = 8/3$.
2. Initial conditions: $[x_1(0), x_2(0), x_3(0)]^T = [0, 5, 75]^T$.
3. Coefficient representation: FLP arithmetic with the number of bits used for the mantissa indicated on each graph.
4. Integration scheme: Forward Euler with step size $\Delta = 1e - 03$.
5. Number of time steps is 25,000.
6. Only projection onto (x_1, x_2)-plane is shown.

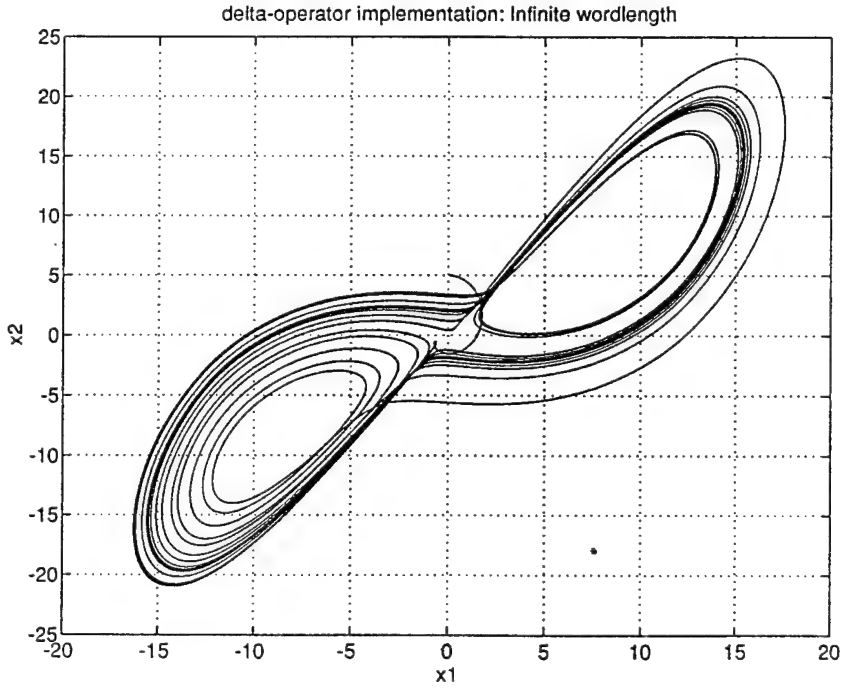
It is important to note that, when only 4 bits are allowed on the mantissa, the qualitative behavior of the q-system is completely different than what is expected. Hoever, the δ-system still provides satisfactory results. Hence, one may use a shorter wordlength for coefficient representation with the latter without affecting performance. The implications on speed, number of components, cost, reliability, etc., are obvious.

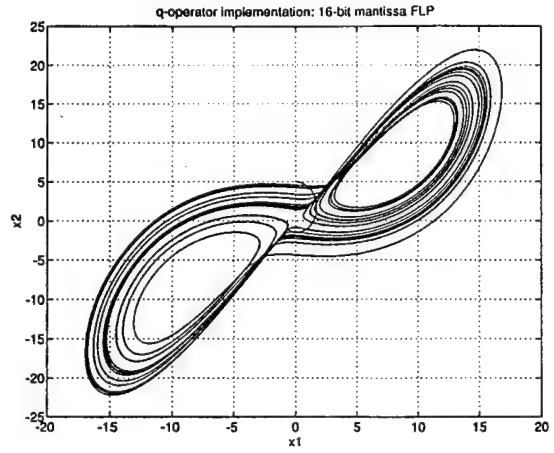
References

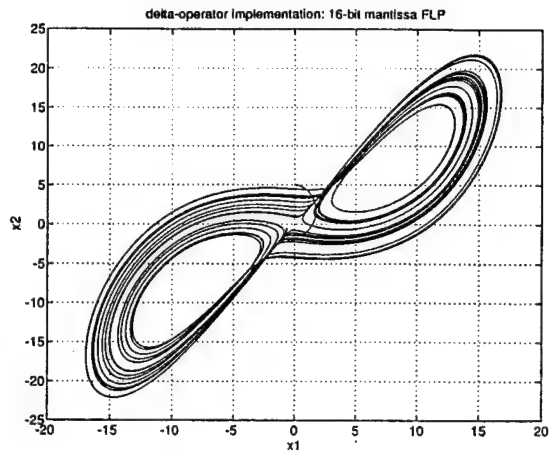
J.W. Brewer (1978). Kronecker products and matrix calculus in system theory. *IEEE Trans. Circ. Syst.,* **CAS-25**, 772-781.

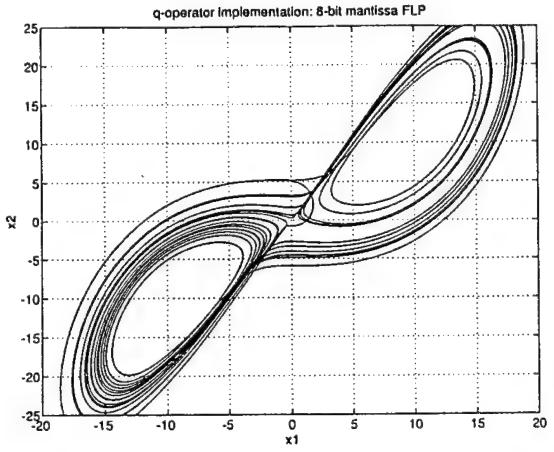
G. Li and M. Gevers (1990). Comparative study of finite wordlength effects in shift and delta operator paramaterization. *Proc. IEEE CDC'90,* Honolulu, HI, **2**, 954-959.

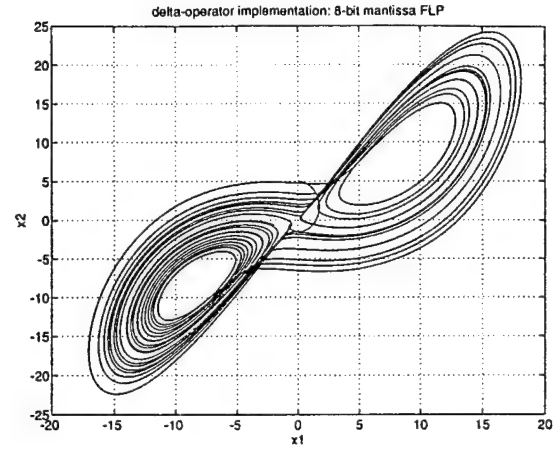

q-operator implementation: Infinite wordlength
x2
x1
25
20
15
10
5
0
-5
-10
-15
-20
-25
-20
-15
-10
-5
0
5
10
15
20

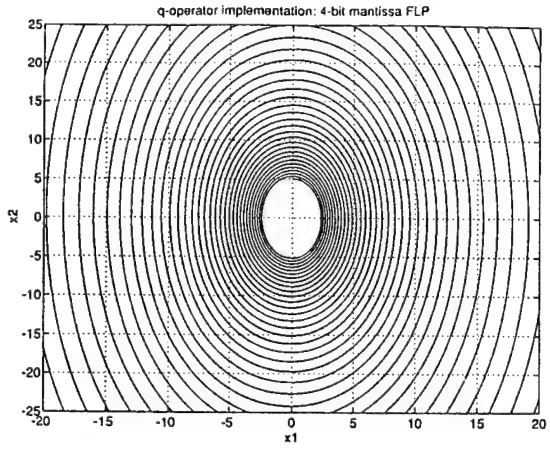

delta-operator implementation: Infinite wordlength
x2
x1
25
20
15
10
5
0
-5
-10
-15
-20
-25
-20
-15
-10
-5
0
5
10
15
20

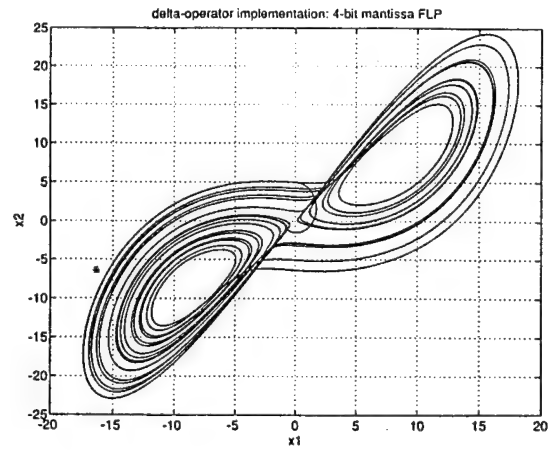

q-operator implementation: 16-bit mantissa FLP
x2
x1


delta-operator implementation: 16-bit mantissa FLP
x2
x1


q-operator implementation: 8-bit mantissa FLP
x2
x1


delta-operator implementation: 8-bit mantissa FLP
x2
x1


q-operator implementation: 4-bit mantissa FLP
x2
x1


delta-operator implementation: 4-bit mantissa FLP
x2
x1

HIGH SPEED FIXED- AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS

Peter H. Bauer and Kamal Premaratne

Focus: Effects of Quantization Errors in Nonlinear Q- and Δ-Operator Systems

Abstract

Absolute quantization error bounds are constructed for q- and δ-operator implementations of the nonlinear system $\underline{x}_{n+1} = f(\underline{x}_n)$. Various assumptions on the type of the nonlinearity $f(\cdot)$ are made and both fixed and floating point formats are investigated. A comparison between the advantages and disadvantages of the two implementation schemes is introduced. Finally, an outlook concerning future work is given.

I. Absolute Bounds on Quantization Errors

I.1. Nonlinearities of the Polynomial Type

I.1.1. Fixed Point Case

Q-operator case:

- **System description:**

$$\underline{x}_{n+1} = f(\underline{x}_n), \quad f(\cdot) : \Re^M \rightarrow \Re^M$$

where

$$f(\underline{x}_n) = \begin{pmatrix} f_1(x_1, \cdots, x_M) \\ \vdots \\ f_M(x_1, \cdots, x_M) \end{pmatrix},$$

$$f_j(x_1, \cdots, x_M) = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} a^{(j)}_{i_1,\cdots,i_M} \, x_1^{i_1} \cdot \cdots \cdot x_M^{i_M}$$

$$j = 1, \cdots, M.$$

- **Assumptions:**
  - **single precision (i.e., single length accumulators)**
  - **quantization step:** q

- **Computed Orbit:**
  $\underline{\hat{x}}(n)$

- **Error model for the computed orbit:**

$$f_j(\hat{x}_1, \cdots, \hat{x}_M) = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} a^{(j)}_{i_1 \cdots i_M} \hat{x}_1^i \cdots \hat{x}_M^{i_M} + \mu_j$$

where

$$\mu_j = \sum_i \epsilon_{ji}^{(1)} + \sum_i \epsilon_{ji}^{(2)} + \cdots + \sum_i \epsilon_{ji}^{(M \cdot N_j)}$$

and

$$|\epsilon_{ji}^{(k)}| < kq \text{ (truncation)}$$
$$|\epsilon_{ji}^{(k)}| \leq \frac{k}{2} q \text{ (rounding)}$$

- **If the nonlinearity $f_j(\cdot)$ is known, the number of nonzero terms in the polynomial is known and therefore the number of terms in the summations of ϵ-terms. Hence an absolute bound on μ_j can be constructed:**

$$|\mu_j| \leq C_j q \text{ (truncation)}$$

where $C_j = l_1 + 2l_2 + \cdots + M \cdot N_j l_{M \cdot N_j}$

$l_\nu, \nu = 1, \cdots, MN_j$ **being the number of terms present in the summation** $\Sigma_i \ \epsilon_{ji}^{(\nu)}$.

δ-operator case:

- **system description:**

$$\delta\underline{x}(n) = \frac{f(\underline{x}_n) - \underline{x}_n}{\Delta}, \quad x(n+1) = x(n) + \Delta\delta[x(n)].$$

$$\delta x_j(n) = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} \frac{a^{(j)}_{i_1 \cdots i_M}}{\Delta} x_1^{i_1} \cdot \cdots \cdot x_M^{i_M} - \frac{x_j}{\Delta} =$$

$$= \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} b^{(j)}_{i_1 \cdots i_M} x_1^{i_1} \cdot \cdots \cdot x_M^{i_M}$$

where

$$b^{(j)}_{i_1,\cdots,i_M} = \begin{cases} \frac{a^{(j)}_{i_1 \cdots i_M} - 1}{\Delta} & \textbf{for } (i_1, \cdots, i_M) \ \ s.th. \\ & i_j = 1, \ and \ i_\nu = 0 \textbf{ for} \\ & \nu = 1, \cdots, M, \ \ \nu \neq j \\ \frac{a^{(j)}_{i_1 \cdots i_M}}{\Delta} & \textbf{otherwise .} \end{cases}$$

- **Assumptions (same as in q-operator case):**
  - **single precision**
  - **quantization step: q**

- **Computed Orbit:**

$$\underline{\hat{x}}(n)$$

- **Error model for the computed orbit:**

$$\delta[\hat{x}_j] = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} b^{(j)}_{i_1 \cdots i_M} \hat{x}_1^{i_1} \cdot \cdots \cdot \hat{x}_M^{i_M} + \mu_j^{(\delta)}$$

where

$$\mu_j^{(\delta)} = \sum_i \mu_{ji}^{(1)} + \sum_i \mu_{ji}^{(2)} + \cdots + \sum_i \mu_{ji}^{(MN_j)}$$

and

$$\begin{aligned} |\mu_{ji}^{(k)}| &< k \cdot q \text{ (truncation)} \\ |\mu_{ji}^{(k)}| &\leq \frac{k}{2} \cdot q \text{ (rounding)} \end{aligned}$$

- **Upper bound on $\mu_j^{(\delta)}$:**

$$|\mu_j^{(\delta)}| \leq (C_j + 1) \cdot q \quad \text{worst case}$$

where C_j is defined as in the q-operator case.

- **Error in the computation of the next state:**

$$\begin{aligned} \hat{x}_j(n+1) &= \hat{x}_j(n) + \Delta\delta[\hat{x}_j(n)] = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} a^{(j)}_{i_1 \cdots i_M} \hat{x}_1^{i_1} \cdots \hat{x}_1^{i_M} \\ &+ \Delta \cdot \mu_j^{(\delta)}(n) + \mu_{\Delta j}^{(\delta)}(n) \end{aligned}$$

where

$$\begin{aligned} |\mu_{\Delta j}^{(\delta)}(n)| &< q \text{ for truncation} \\ |\mu_{\Delta j}^{(\delta)}(n)| &< \frac{q}{2} \text{ for rounding} \end{aligned}$$

Comparison: δ- vs. q-operator:

Error term bound for the q-operator:

$$|\mu_j| \leq C_j \cdot q \text{ (truncation)}$$

Error term bound for the δ-operator:

$$|\Delta \cdot \mu_j^{(\delta)} + \mu_{\Delta j}^{(\delta)}| \leq (C_j+1)q\Delta + q = q([C_j+1]\Delta+1) \text{ (truncation)}$$

- **δ-operator formulation has a smaller absolute error bound for:**

$$\frac{C_j - 1}{C_j + 1} > \Delta, \text{ for } j = 1, \cdots, M$$

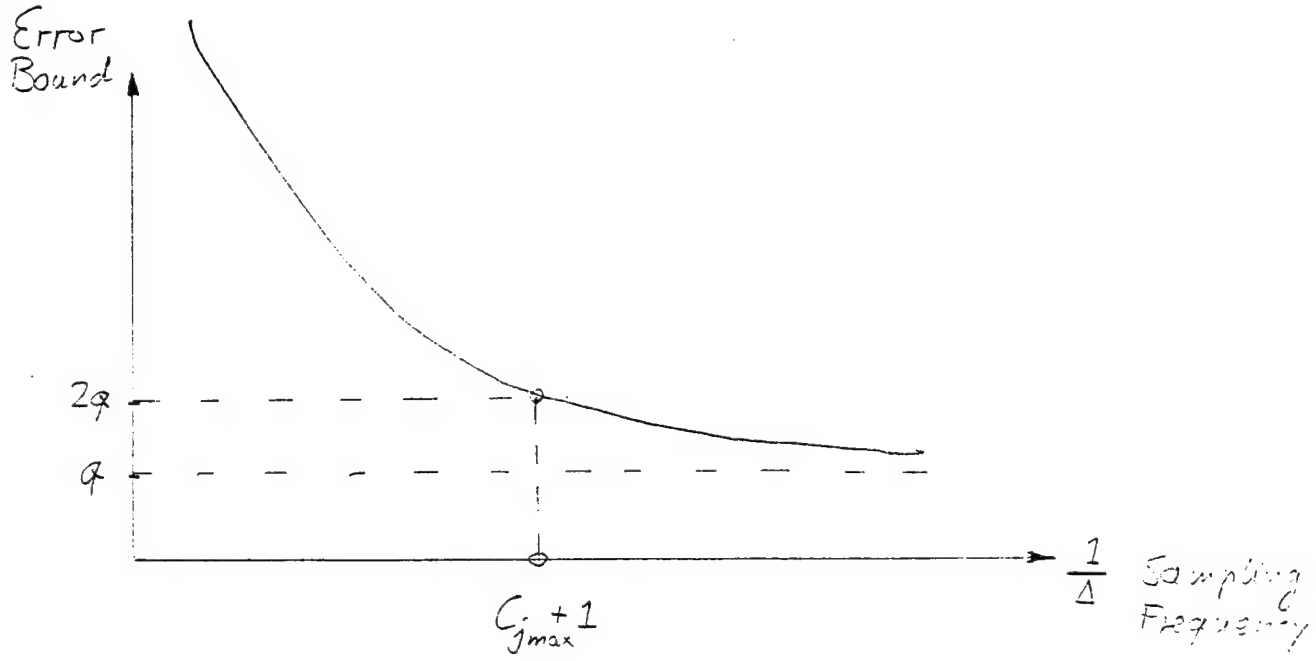
Usually $C_j >> 1$ and hence for

$$1 - \epsilon > \Delta,$$

the δ-operator system is preferable.
(For high speed systems we typically have $\Delta << 1$.)

Reasonable choice of Δ (from an error bound perspective):

$$\Delta \leq \frac{1}{C_{j\max} + 1}$$



Remarks:

- δ-operator implementations in FXP format seem to produce a significantly smaller bound than q-operator implementations, if $\Delta << 1$.
- A δ-operator implementation requires a larger dynamic range than the q-operator implementation, if $\Delta << 1$. $\Rightarrow$ the chance of overflow increases.
- To avoid overflow problems, the δ-operator system needs to be implemented with a larger wordlength.

Forced Response Case:

System description for the q-operator:

$$\begin{aligned} \underline{x}_{n+1} &= f(\underline{x}_n, \underline{u}_n) \\ f(\underline{x}_n, \underline{u}_n) &= \begin{pmatrix} f_1(x_1, \cdots, x_M, u_1, \cdots, u_K) \\ \vdots \\ f_M(x_1, \cdots, x_M, u_1, \cdots, u_K) \end{pmatrix} \end{aligned}$$

where the f_ν's are again multivariate polynomials in up to $M + K$ variables.

System description for the δ-operator:

$$\begin{aligned} \delta[\underline{x}_n] &= \frac{f(\underline{x}_n) - \underline{x}_n}{\Delta} \\ \underline{x}_{n+1} &= \underline{x}_n + \Delta\delta[\underline{x}_n] \end{aligned}$$

- **Using a similar error model as in the zero-input case, the computation of $\underline{x}_{n+1}$ in the q-operator case and the computation of $\delta\underline{x}_n$ in the δ-operator case produce bounds of similar magnitude.**
- **If $\Delta << 1$, the δ-operator system again has an advantage over the q-operator system, since the errors of the first equation are much larger than the ones produced in the update equation.**

I.1.2. The Floating Point Case

Q-0perator model – ideal case:

$$x_j(n+1) = \sum_{i_1} \cdots \sum_{i_M} a^{(j)}_{i_1 \cdots i_M} \, x_1^{i_1} \cdots x_M^{i_M} \tag{1}$$

δ-operator model – ideal case:

$$\delta x_j(n) = \sum_{i_1} \cdots \sum_{i_M} \frac{a^{(j)}_{i_1 \cdots i_M}}{\Delta} x_1^{i_1} \cdots x_M^{i_M} - \frac{x_j}{\Delta} \tag{2a}$$

$$x_j(n+1) = x_j(n) + \Delta \delta[x_j(n)] \tag{2b}$$

Model for floating point errors due to multiplication and addition:

$$\begin{aligned} x \odot y &= xy(1+\epsilon) \\ x \oplus y &= x(1+\epsilon) + y(1+\epsilon_2) \end{aligned}$$

Consider two cases:

(a) $a_{i_1,\cdots,i_M} \simeq 1$ **with** $i_\nu = 0$ **for** $\nu = 1, \cdots, M, \nu \neq j$ **and** $i_j = 1$.
$|a_{i_1,\cdots,i_M}| << 1$ **for all other combinations of** $(i_1, \cdots, i_M)$,
$\underline{x}(n) \epsilon [-1, +1]^M$

(b) condition (a) is not satisfied.

Case (a):

δ-operator bounds on quantization error are much smaller than q-operator bounds.

Qualitative explanation:

For case (a), all partial sums and products in the computation of $\Delta \cdot \delta[x_j(n)]$ are much smaller than $x_j(n)$. Therefore the errors in the computation of $\Delta\delta[x_j(n)]$ are smaller compared to the final addition error in (2b). Therefore the δ-operator model implicitely performs operand sorting, which is known to reduce quantization errors in floating point arithmetic.

Case (b):

δ-operator error bounds are slightly larger than q-operator bounds.

Other classes of nonlinear systems exist which also perform better using a δ-operator formulation. One such class is the weakly nonlinear functions satisfying:

$$\begin{aligned} f_j(x_1, \cdots, x_M) &= x_j + \epsilon_j(x_1, \cdots x_M) \\ &\quad \text{with } |\epsilon_j(x_1, \cdots, x_M)| << |x_j| \\ &\quad j = 1, \cdots, M. \end{aligned}$$

(If $f_j(x_1, \cdots, x_M)$ is of polynomial type, the system has to operate on a hypercuboid or another finite subspace of $\Re^M$ since polynomials cannot be weakly nonlinear in the above sense for all $x_i \epsilon \Re$.)

Note:

If equations (1) or (2) arise from quantizing a continuous time system with a very short sampling time, then the condition

$$|x_j(n)| >> |\Delta\delta x_j(n)|$$

can be satisfied giving the δ-operator formulation an advantage over the q-operator.

I.1.3. A Generalized Delta-Operator Model for Linear and Nonlinear Systems

Linear Case:

Assume the system is given by:

$$\underline{x}(n+1) = A_q \underline{x}(n) + B_q \underline{u}(n)$$

Consider the modified Δ-operator form:

$$\begin{aligned} \delta[x(n)] &= \frac{A_q - A_0}{\Delta}\underline{x}(n) + \frac{B_q - B_0}{\Delta}\underline{u}(n) \\ x(n+1) &= A_0\underline{x}(n) + B_0\underline{u}(n) + \Delta\delta[x(n)] \end{aligned}$$

with A_0 and B_0 being integer matrices closest to A_q and B_q respectively.

Advantages

- **The dynamic range of the A^δ and B^δ matrices becomes smaller and the chance of overflow is reduced.**
- **The delta operator realization has the same improved sensitivity as in the regular delta-operator case.**
- **In floating point arithmetic, the condition $A_q \simeq I$ is not necessary for improved error behavior of the delta-operator system.**

Nonlinear Case:

A similar argument as in the linear case can be made for weakly nonlinear systems of the form:

$$\underline{x}(n+1) = A_q \underline{x}(n) + B_q \underline{u}(n) + \underline{\epsilon}(\underline{x}(n), \underline{u}(n))$$

where

$$\| \underline{\epsilon}(\underline{x}(n), \underline{u}(n)) \| << \| \underline{x}(n) \|$$

and $$\| \underline{\epsilon}(\underline{x}(n), \underline{u}(n)) \| << \| \underline{u}(n) \|$$

I.2. Nonlinearities of Piecewise Linear Form

I.2.1. The Fixed Point Case

Although a piecewise linear continuous scalar function f: $\Re \rightarrow \Re$ can be represented as

$$f(x) = \sum_i (|x - \mu_i| a_i) + b,$$

a computationally more efficient realization is:

$$f(x) = c_i x + d_i \text{ for } \underline{x}_i \leq x \leq \bar{x}_i \tag{1}$$

Therefore, the resulting system $\underline{x}_{n+1} = f(\underline{x}_n)$ can be written in form of several linear state space equations with a driving term, and the driving terms being known *a priori*:

δ – operator:

$$\delta[\underline{x}_n] = A_i^\delta \underline{x}_n + \underline{u}_i^\delta$$
$$\underline{x}_{n+1} = \underline{x}_n + \Delta\delta[\underline{x}_n],\ i = 1, \cdots, K$$

q – operator:

$$\underline{x}_{n+1} = A_i^q \underline{x}_n + \underline{u}_i^q,\ i = 1, \cdots, K$$

Conclusion:

- For single precision (quantization after products) the absolute error bounds for the δ-operator realization are smaller than for the q-operator realization.
- For double precision (quantization after summation) the absolute error bounds for the δ-operator realization are approximately the same as for the q-operator.

I.2.2. The Floating Point Case

Due to (1), the system can be modeled as a time-variant linear system with a known, piecewise constant input. Therefore the same conclusions apply as in the linear t.i.v. case with regard to absolute error bounds:

- Generally, absolute error bounds of the δ- and q-operator system are of similar size.

- If the resulting A-matrices of the piecewise linear system are all 'close' to the identity matrix I, then the δ-operator system will perform superior to the q-operator. (see comments in I.1.2.). This requires that the driving terms are also small relative to the states.

I.3. Sector Bounded Nonlinear Functions

I.3.1. The Fixed Point Case

System description:

$$\begin{aligned} x_i(n+1) &= \mathcal{F}_{i1}[a_{i1}x_1(n)] + \cdots + \mathcal{F}_{im}[a_{im}x_m(n)] \\ i &= 1, \cdots, m. \end{aligned}$$

Sector conditions on $\mathcal{F}[\ ]$:

$$\mathcal{F}_{ij}(x) = k_{ij}x,\ k_{ij} \in [\underline{k}_{ij}, \bar{k}_{ij}].$$

If $\epsilon_{ij}(n)$ is the error affiliated with the computation of $\mathcal{F}_{ij}[\cdot]$, the response of the q and δ-operator system can be absolutely bounded by the following majorant system:

q-operator:

$$x_i^+(n+1) = \sum_{j=1}^{m} m_{ij}^+ x_j^+(n) + \sum_{j=1}^{m} \epsilon_{ij}^+(n),\ \ i = 1, \cdots, m$$

where

$$\begin{aligned} m_{ij}^+ &= \max\{|\underline{k}_{ij}a_{ij}|, |\bar{k}_{ij}a_{ij}|\} \\ \epsilon_{ij}^+(n) &= |\epsilon_{ij}(n)| \end{aligned}$$

δ-operator

$$x_i^+(n+1) = \sum_{j=1}^{m} m_{ij}^+ x_j^+(n) + \Delta\left(\epsilon_\Delta^+(n) + \sum_{j=1}^{m} \epsilon_{ij}^+(n)\right) + \epsilon_{up}^+(n),$$
$$i = 1, \cdots, m$$

where

$\epsilon_{up}^+(n) = |\epsilon_{up}(n)|$, $\epsilon_{up}(n)$: **error occurring in update equation**
$\epsilon_\Delta^+(n) = |\epsilon_\Delta(n)|$, $\epsilon_\Delta(n)$: **error due to division by Δ.**

Comparison:

The bound for the δ-operator implementation is lower if

$$\max\left(\Delta(\epsilon_\Delta^+(n) + \sum_{j=1}^{m} \epsilon_{ij}^+(n)) + \epsilon_{up}^+(n)\right) < \max\left(\sum_{j=1}^{m} \epsilon_{ij}^+(n)\right)$$

- **Since the bound for $|\epsilon_{ij}^+(n)|$ is typically much larger than for $|\epsilon_\Delta^+(n)|$ or $|\epsilon_{up}^+(n)|$, it is obvious that for $\Delta << 1$, the above condition is always satisfied.**
- **A similar result holds if the nonlinearities $\mathcal{F}$ enter the system in a different form, i.e., if they have arguments which consist of partial sums.**
- **A similar comparison arises for other fixed point quantization formats.**

I.3.2. The Floating Point Case

Generally, the δ-operator implementation is not superior to the q-operator implementation if one compares absolute error bounds. However, as stated before (I.1.2, I.2.2), if the condition

$$|x_i(n)| >> |\Delta\delta(x_i(n))|$$

holds true for all states $(i = 1, \cdots, m)$, then the δ-operator implementation has a significantly smaller error bound.

A class of systems satisfying the above condition is given by:

$$\begin{aligned} x_i(n+1) &= \mathcal{F}_{i1}[a_{i1}x_1(n)] + \cdots + \mathcal{F}_{im}[a_{im}x_m(n)] \\ & \quad i = 1, \cdots, m \end{aligned}$$

where

$$\begin{aligned} \mathcal{F}_{ij}(x) &= k_{ij}x, \quad k_{ij} \in [\underline{\epsilon}_{ij}, \overline{\epsilon}_{ij}] \text{ for } i \neq j, \\ & \qquad k_{ii} \in [1 - \epsilon_{ii}, 1 + \epsilon_{ii}] \text{ otherwise,} \\ & \quad |\epsilon_{ij}| << 1 \text{ for } i, j = 1, \cdots, m. \end{aligned}$$

Again, such a system could arise from a continuous time system with a high sampling rate.

II. Comparison of Implementations

	FXP-case	FLP-case	BFLP-case*
general system: q-error bounds	δ-operator is mostly superior	δ and q-operator are comparable	similar to FXP case?
q-error bounds for a short sampling time in the discretization process	δ-operator is mostly superior	δ operator is superior	similar to FXP case?
limit cycles (linear case only)	δ-operator produces incorrect equilibria	limit cycles in underflow for both q and δ-operator	similar to FLP case
hardware requirements for small Δ	δ-operator requires longer registers than q-operators	independent of Δ	
overflow effects	δ-operator is more likely to cause overflow	in both operators unlikely	similar to FLP case
general sensitivity	δ operator superior	δ-operator better than or equal to q-operator	similar to FXP case?
sensitivity for a short sampling time in discretization process	δ-operator superior	δ operator superior	similar to FXP case?

* has not been analyzed in detail yet, expected results.

SEMIANNUAL PERFORMANCE REPORT
GRANT NO's: N00014-94-1-0387

Summary of Phase P1 Results

Phase P1 consists of two tasks:

[T1] Task T1: Analysis and design of finite wordlength implementations of linear, time-invariant δ-Systems.

[T3] Task T3: 2-D and m-D δ-system models.

The major part of task T1 was carried out at the University of Notre Dame by Dr. Peter H. Bauer while the major part of task T3 was carried out at the University of Miami by Dr. Kamal Premaratne under grant No. N00014-94-1-0454. The project being an extensive collaborative effort, the two PI's have been in constant contact during this research effort.

The following is a summary of the phase P1 results.

Task T1: Analysis and Design of Finite Wordlength Implementations of Linear, Time-Invariant δ-Systems

The conclusions drawn from the work conducted for task T1 may be summarized as follows:

1. The Fixed-Point Arithmetic Case: When limit cycle performance is crucial, the q-operator implementation is preferrable. The δ-operator implementation is superior with regard to coefficient sensitivity issues.

2. The Floating-Point Arithmetic Case: Generally, the δ-operator implementation outperforms its q-operator counterpart. In particular, in high-order and high-speed applications, the δ-operator implementation is the best choice.

Prior to a more detailed exposition, first we provide qualitative justification for the above conclusion. The state equations of a δ-operator system can be written as:

$$\begin{aligned} \delta[\mathbf{x}](n) &= A_\delta \mathbf{x}(n) + B_\delta \mathbf{u}(n); \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n). \end{aligned} \tag{T1.1}$$

where $\mathbf{x}$ and $\mathbf{u}$ are the state and input vectors, respectively. Here, Δ denote a positive real

constant (typically, the sampling time). The symbol $\delta[\cdot]$ denotes the δ-operator, that is,

$$\delta[\mathbf{x}](n) = \frac{q[\mathbf{x}](n) - \mathbf{x}(n)}{\Delta} = \frac{q-1}{\Delta}\mathbf{x}(n), \tag{T1.2}$$

and $q[\cdot]$ denotes the usual q-operator, that is,

$$q[\mathbf{x}](n) = \mathbf{x}(n+1). \tag{T1.3}$$

The corresponding formulation of (T1.1) in terms of the q-operator is

$$q[\mathbf{x}](n) = A_q\mathbf{x}(n) + B_q\mathbf{u}(n), \tag{T1.4}$$

where

$$A_q = I + \Delta \cdot A_\delta \Longleftrightarrow A_\delta = \frac{A_q - I}{\Delta} \quad \text{and} \quad B_q = \Delta \cdot B_\delta \Longleftrightarrow B_\delta = \frac{B_q}{\Delta}. \tag{T1.5}$$

Now, given $\mathbf{x}$ and $\mathbf{u}$, both representations compute $q[\mathbf{x}]$ with a certain accuracy. Consider the δ-operator formulation in (T1.1). Here we encounter two errors:

1. The first is due to the computation of $\delta[\mathbf{x}]$, that is, the first equation in (T1.1). We will refer to this equation as the *intermediate equation.*

2. The second is due to the eventual computation of $q[\mathbf{x}]$, that is, the second equation in (T1.1). We will refer to this equation as the *update equation.*

Let us assume that the total error in computing $q[\mathbf{x}]$ is mainly due to the intermediate equation in (T1.1) (rather than the update equation). Then, by choosing Δ sufficiently small, the total error in computing $q[\mathbf{x}]$ will be approximately the error created by the update equation which is small!. In this case, the δ-operator representation has better finite wordlength properties than its q-operator counterpart in (T1.4).

If, however, the errors accumulated in the intermediate and the update equations in (T1.1) are comparable, $q[\mathbf{x}]$ computed through the δ-operator representation will show approximately the same error as that computed through its q-operator counterpart assuming Δ is sufficiently small. If Δ is not sufficiently smaller than one, the δ-operator representation will actually perform worse than the q-operator representation!

If the error introduced in the update equation is larger than that in the intermediate equation, the δ-operator representation would consistently perform worse!! In reality, this case is very unlikely to occur.

Next, a more detailed exposition follows.

T1.1 The Fixed-Point Arithmetic Case

We now discuss some of the results regarding the fixed-point (FXP) case. Here, our results in fact indicate that, in case limit cycle behavior is crucial, the δ-operator representation is NOT suitable with this arithmetic scheme [1]. Such a case may occur when nonlinear systems are implemented through FXP δ-operator based schemes.

Zero-input limit cycles. Independent of Δ, zero-input limit cycles cannot be avoided in FXP δ-implementations. This is easily explained as follows: If Δ is chosen very small, the contribution from the intermediate equation being small (since $\delta[\mathbf{x}]$ is being multiplied by Δ), during the update equation, $q[\mathbf{x}]$ can be quantized to $\mathbf{x}$ creating a DC limit cycle, that is, an incorrect equilibrium point different from zero results. We emphasize that, most of the desirable properties of δ-operator implementations are based on a small Δ. We may also show that, if Δ is chosen larger (this case is of course somewhat less important), DC limit cycles will still exist. Hence, δ-operator representations cannot be implemented limit cycle free in FXP format! This fact is independent of the particular realization of the system.

Deadband size. Since δ-systems cannot be implemented limit cycle free in FXP format, it is of interest to investigate te the size of such limit cycles since, in certain situations, such small limit cycle amplitudes can be tolerated. It can be shown that, the magnitude of Δ determines the magnitude of the limit cycle. The smaller the Δ, the larger will be the deadband and hence the limit cycle magnitude. An approximate relationship regarding this is

$$\Delta \times \text{size of deadband} = 1, \tag{T1.6}$$

where the size of deadband is measured in multiples of the quantization step size. Here, the deadband corresponds to that obtained by considering the quantization of $\Delta \cdot \delta[\mathbf{x}]$. Therefore, the usual choice of a small Δ creates a larger deadband!

The input driven case. Although the input driven case is not part of the originally proposed work, some interesting results have been obtained. For small values of Δ, there exists a bounded input signal that does not allow control of the state trajectory. In other words, given sufficiently small Δ, the state trajectory may not be influenced by such an input signal.

The influence of the realization. First, it was necessary to develop a suitable scheme

to investigate the effect of realization on the presence or absence of limit cycles. In this direction, for the q-operator case, a computer-based exhaustive search algorithm that checks for limit cycles (DC and/or oscillatory) has been developed [5].

As discussed before, we have shown that, a stable linear time-invariant δ-system cannot be implemented limit cycle free in FXP. The size of the deadband however also depends on the particular realization, that is, the structure of A_δ. Given a system transfer function, there are forms which minimize this deadband size with respect to some appropriately chosen measure. For example, in order to minimize DC limit cycle amplitude, one may choose the normal form (in terms of A_δ) as a suitable candidate.

The influence of quantization nonlinearity and its deadzone. Since a larger deadzone implies larger DC limit cycle amplitudes, the use of quantizers with reduced, or even zero, deadzone was therefore proposed. In investigating first-order systems, by reducing the deadzone, it was found that, existence of DC limit cycles can indeed be reduced. Unfortunately, other oscillatory limit cycles will be created. This phenomenon is due to the increased gain exhibited towards small input signals by the quantizer.

Scaling. As discussed above, we have shown that, independent of either the form of A_δ or the magnitude of Δ, a FXP implemented δ-system cannot be free of zero-input limit cycles. Hence, scaling cannot be offered as a possible solution.

T1.2 The Floating-Point Arithmetic Case

The floating-point (FLP) implementation of δ-systems is currently under investigation. The results obtained so far are very encouraging, and indicate that, quantization errors due to FLP arithmetic have a much smaller effect on the system behavior than in the FXP case. In fact, preliminary results show that, for δ-systems of order three and higher, errors in computing $q[\mathbf{x}]$ can be made significantly smaller than for the corresponding q-systems. This is because, for a FLP implementation of such a system, errors created through the intermediate equation are larger than those created through the update equation. As previously mentioned, in this situation, δ-systems behave better than their q-operator counterparts!

Limit cycles. In FLP arithmetic, a linearly stable time invariant system, under zero-input conditions, may exhibit four types of responses: A diverging response, an oscillatory periodic response of arbitrary magnitude, an oscillatory periodic response in underflow, or an asymptotically stable response. Only the last two response types are acceptable in practice. It is well known that, the last response type is in fact a very stringent requirement

and is often not required in practice. Results so far obtained show that, when the requirements for a response in underflow are compared, the δ-system requires less wordlength than its q-system counterpart! This advantage in fact grows with the order of the system!!

Once the system reaches underflow conditions, the δ-system again exhibits DC limit cycles. However, if the exponent register is chosen sufficiently large, the amplitude of these oscillations can be made extremely small and hence, for all practical purposes, this problem is solved.

Deadband size. If the condition on the mantissa length that guarantees convergence into underflow is satisfied, then the deadband size will be very small. Hence, it can be neglected for all practical purposes. This assumes a properly chosen exponent register length since the exponent register length determines the dynamic range of underflow.

The Influence of the Nonlinearity. Unlike the FXP case, the characteristic of the nonlinearity has only a minor effect on the system behavior, significant differences being present only in underflow conditions

The Underflow case. In underflow, the δ-system seems to behave worse than its q-operator counterpart. This is mainly due to the fact that, a FLP system in underflow essentially performs very similar to a FXP system. However, as mentioned above, if the dynamic range of underflow is chosen properly, the system behavior in underflow is of little practical interest.

Block Floating-Point Arithmetic. Even for the q-operator case, results regarding block FLP implementations are lacking. Hence, investigations regarding block FLP implementation of δ-systems is in its early stages. In order to obtain a comparison between the two types of implementations, current research is geared towards obtaining results applicable for the q-operator case.

T1.3 The Multi-Dimensional Case

The results on one-dimensional (1-D) δ-operator implementations in FXP arithmetic directly carry over to the multi-dimensional (m-D) case. The existence of non-converging responses along the boundary of the causality region can easily be proven using the same type of argument used in the 1-D case. Consequently, δ-operator based implementations of m-D systems cannot be implemented limit cycle free in FXP.

Task T3: 2-D and m-D δ-system models

Discrete-time systems implemented using the δ-operator, as is clear from the discussion above, exhibit superior finite wordlength properties with FLP arithmetic. In the case of FXP arithmetic, they still provide superior coefficient sensitivity. The development of 2-D and m-D models applicable for δ-operator implementations was hence motivated with the expectation that these properties would still hold true.

The conclusions drawn from the work conducted for task T3 may be summarized as follows: Similar to the 1-D case, under FLP arithmetic, the δ-operator implementation of 2-D and m-D discrete-time systems provides the best choice. Again, this is particularly true in high-order and high-speed applications.

State-space models. In Roesser local s.s. model of q-operator formulated 2-D discrete-time systems takes the form

$$\begin{aligned}
\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [A_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_q]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j) \\
&\doteq [C_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j),
\end{aligned} \tag{T3.1}$$

where $A_q^{(1)}$ is of size $n_h \times n_h$, $A_q^{(4)}$ is of size $n_v \times n_v$, etc. Also, $q_h[\cdot]$ and $q_v[\cdot]$ denote the horizontal and vertical shift operators, that is,

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{T3.2}$$

To exploit the advantages of δ-operator implementations, analogous to the 1-D case, we define the operators

$$\begin{aligned}
\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\Delta_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h}; \\
\delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\Delta_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v},
\end{aligned} \tag{T3.3}$$

where Δ_h and Δ_v are two positive real constants. The corresponding δ-operator s.s. model

may then be obtained as

$$\begin{aligned}
\begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [A] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= [\,C^{(1)} \quad C^{(2)}\,] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j) \\
&\doteq [C] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j).
\end{aligned} \tag{T3.4}$$

This is the 2-D version of the intermediate equation mentioned earlier. In addition, as for the 1-D case, we have the following update equations:

$$\begin{aligned}
q_h[\mathbf{x}^h](i,j) &= \mathbf{x}^h(i,j) + \Delta_h \cdot \delta_h[\mathbf{x}^h](i,j); \\
q_v[\mathbf{x}^v](i,j) &= \mathbf{x}^v(i,j) + \Delta_v \cdot \delta_v[\mathbf{x}^v](i,j).
\end{aligned} \tag{T3.5}$$

Note that,

$$\begin{aligned}
A_q = I + \boldsymbol{\Delta} \cdot A_\delta &\iff A_\delta = \boldsymbol{\Delta}^{-1} \cdot (A_q - I_n); \\
B_q = \boldsymbol{\Delta} \cdot B &\iff B_\delta = \boldsymbol{\Delta}^{-1} \cdot B_q; \\
C_q = C_\delta &\iff C_\delta = C_q; \\
D_q = D_\delta &\iff D_\delta = D_q.
\end{aligned} \tag{T3.6}$$

Here, $\boldsymbol{\Delta} = [\Delta_h I_{n_h} \oplus \Delta_v I_{n_v}]$ is of size $(n_h + n_v) \times (n_h + n_v)$.

The associated system theoretic notions, such as, transition matrix, transfer function, characteristic equation, etc., have also been introduced. This s.s. model is the basis for designing 2-D filters with superior finite wordlength properties. The design procedures developed are expected to be extremely useful in obtaining high-Q 2-D and m-D digital filters that are suitable for high-speed applications.

Stability. In the 1-D case, it has been shown that, direct techniques with no recourse to transformations (that first converts a given δ-system to its q-system counterpart) can provide numerically more reliable stability checking algorithms. With this in mind, for the 2-D case, a direct stability checking technique applicable to the corresponding δ-system transfer function has been introduced. For this purpose, a recently developed tabular form was extended to the complex coefficient case and the notion of Schur-Cohn minors was introduced to the δ-operator case.

Gramians and balanced realization. The notions of reachability and observability gramians and balanced realization have been introduced for the δ-operator case. In order

to do this, first, the relationship between the gramians for the δ- and q-operator cases, as defined in the literature, was established. The reachability and controllability gramians, that is, P and Q, respectively, for 1-D δ-systems were found to satisfy

$$\begin{aligned} P &= \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} (cI - A_\delta)^{-1} B_\delta B_\delta^* (c^* I - A_\delta^*)^{-1} \frac{dc}{1 + \Delta c}; \\ Q &= \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} (c^* I - A_\delta^*)^{-1} C_\delta^* C_\delta (cI - A_\delta)^{-1} \frac{dc}{1 + \Delta c}, \end{aligned} \tag{T3.7}$$

where $\mathcal{T}_\delta$ is the stability boundary applicable for δ-systems, that is, $\mathcal{T}_\delta = \{c \in \Im : |c + 1/\Delta| = 1/\Delta\}$. An extension of this is then used to define the 2-D gramians of δ-systems represented in the Roesser model developed above.

For the important class of separable (that is, separable-in-denominator) systems, it is shown that these gramians may be computed through the solution of four Lyapunov equations. These notions and results are useful in many applications, such as, in extracting reduced order models of δ-systems.

Sensitivity. Measures that indicate coefficient sensitivity of the δ-models developed above have been introduced. Unlike what is available in literature, this development is applicable to the MIMO case as well. With these sensitivity measures as a guide, development of minimum sensitivity structures has been carried out. The connection with the corresponding balanced realizations has been pointed out.

Roundoff noise. With the use of a noise model that takes into account the roundoff error propagation in the s.s. model developed above, structures that minimize roundoff noise have been developed.

<u>Publications:</u> Work directly related to grants

[1] K. Premaratne and P.H. Bauer (1994). Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic. *Proceedings 1994 IEEE International Symposium on Circuits and Systems (ISCAS'94),* London, UK, vol. 2, 461-464.

[2] P.H. Bauer and K. Premaratne (1994). Fixed-point implementation of multi-dimensional delta-operator formulated discrete-time systems: Difficulties in convergence. *Proceedings of the 1994 IEEE SOUTHEASTCON,* Miami, FL, 26-29.

[3] K. Premaratne and A.S. Boujarwah (1994). An algorithm for stability determination

of two-dimensional delta-operator formulated discrete-time systems. *Multidimensional Systems and Signal Processing,* to appear.

[4] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer (1994). Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its finite wordlength properties. *37th Midwest Symposium on Circuits and Systems,* Lafayette, LA, to be presented; *IEEE Transactions on Signal Processing,* in preparation.

[5] E.C. Kulasekere, K. Premaratne, P.H. Bauer, and L.J. Leclerc (1994). An exhaustive search algorithm for checking limit cycle behavior of digital filters. *IEEE Transactions on Signal Processing,* in preparation.

Note. The contents of [1] and [2] are also being prepared for possible publication in *IEEE Transactions on Signal Processing.*

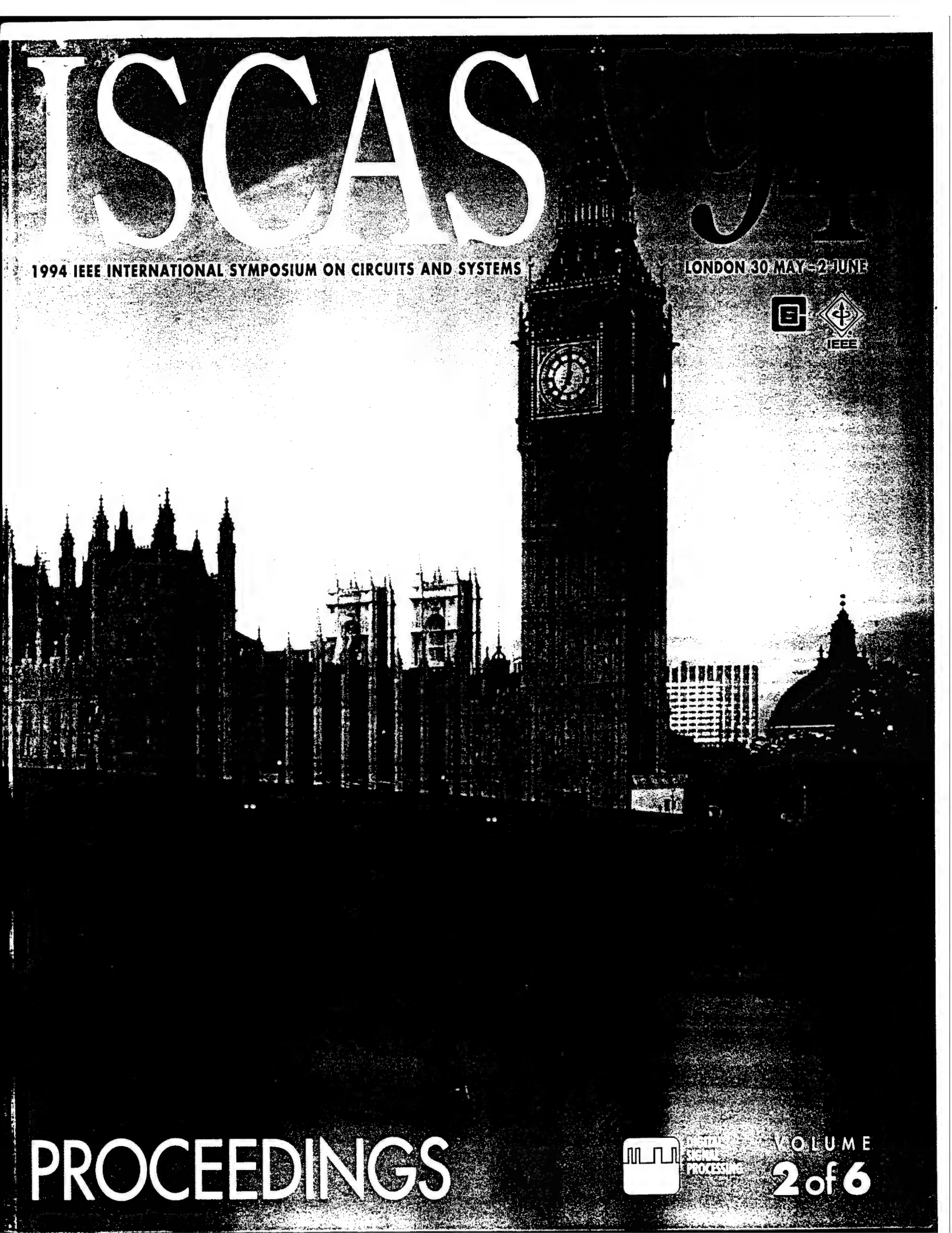
Publications: Other work where grants are acknowledged

[1] K. Premaratne and E.I. Jury (1994). Discrete-time positive-real lemma revisited: The discrete-time counterpart of the Kalman-Yakubovitch lemma. *IEEE Transactions on Circuits and Systems—I. Fundamental Theory and Applications,* to appear.

[2] M.M. Ekanayake and K. Premaratne (1994). Two-channel IIR QMF filter banks with approximately linear-phase analysis and synthesis filters. *28th Annual Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, to be presented; *IEEE Transactions on Signal Processing,* in review.

[3] K. Premaratne and M. Mansour (1994). Robust stability of time-variant discrete-time systems with bounded parameter perturbations. *IEEE Transactions on Circuits and Systems—I. Fundamental Theory and Applications,* in review.

[4] S.A. Yost and P.H. Bauer (1994). Robust stability of multi-dimensional difference equations with shift-variant coefficients. *Multidimensional Systems and Signal Processing,* to appear.


ISCAS
1994 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS
LONDON 30 MAY - 2 JUNE
IEEE
PROCEEDINGS
VOLUME
2 of 6

Limit cycles and asymptotic stability of delta-operator formulated discrete-time systems implemented in fixed-point arithmetic

Kamal Premaratne
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124
USA
(+1) 305-284-4051
kprema@umiami.ir.miami.edu

Peter H. Bauer
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556
USA
(+1) 219-631-8015
pbauer@mars.ee.nd.edu

ABSTRACT

This paper analyzes the problem of global asymptotic stability of delta-operator formulated discrete-time systems implemented in fixed-point arithmetic. It is shown that the free response of such a system tends to produce period one limit cycles if conventional quantization arithmetic schemes are used. Explicit necessary conditions for global asymptotic stability are derived, and these demonstrate that, in almost all cases, fixed-point arithmetic does not allow for global asymptotic stability in delta-operator formulated discrete-time systems that use a short sampling time.

I. INTRODUCTION

Recently, discrete-time systems formulated with the incremental difference operator (or, δ-operator) have been receiving considerable attention in the technical literature [1-4]. Most of this work focus on its superior performance under finite wordlength conditions when compared with those formulated with the shift-operator (or, q-operator). In particular, investigations of coefficient sensitivity and quantization noise properties have revealed that δ-operator formulations usually perform significantly better than their q-operator counterparts [1-4]. This is especially true for high-speed applications where the sampling rate is much larger than the underlying system bandwidth. Under these conditions, q-operator formulated discrete-time systems tend to become ill-conditioned [1-2].

Although a large amount of work is available on the effects of coefficientsensitivity and quantization noise, a deterministic study of the nonlinear behavior of discrete-time systems formulated with the δ-operator has not been undertaken. In the case of floating-point (FLP) arithmetic, some results for feedback system are available in [2].

In this work, we focus on the convergence behavior of the unforced system response and global asymptotic stability of δ-operator formulated discrete-time systems implemented in fixed-point (FXP) arithmetic. In particular, via necessary conditions for stability, it will be shown that such systems tend to produce DC limit cycles.

The structure of this article is as follows: In Section II, we introduce notation and nomenclature. The model for δ-operator formulated discrete-time systems, with and without quantization nonlinearities, is briefly discussed. Section III addresses the problem of asymptotic stability when FXP arithmetic is used for the implementation. In terms of ensuing DC limit cycles, necessary conditions for global asymptotic stability are formulated. It is shown that, when FXP arithmetic is used, stability of the linear system is often lost. Section IV provides concluding remarks.

II. NOTATION AND NOMENCLATURE

Since our focus is on investigation of stability properties of δ-operator formulated discrete-time systems under unforced conditions, the state equations of the system under zero-input will be considered.

In the linear case, the general m-th order state-space representation is given by

$$\delta[\mathbf{x}](n) = A^{\delta}\mathbf{x}(n); \tag{1}$$

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n), \tag{2}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_m(n)]^T$ is the state vector at instant n, $A^{\delta} = \{a_{ij}^{\delta}\} \in \Re^{m \times m}$ is the system matrix,

and $\Delta > 0$ is the sampling time. Moreover, $\delta[\cdot]$ represents the δ-operator, that is,

$$\delta[x_\nu](n) = \frac{x_\nu(n+1) - x_\nu(n)}{\Delta}, \ \forall \nu = 1, \ldots, m, \quad (3)$$

and $\delta[\mathbf{x}](n) = [\delta[x_1](n), \ldots, \delta[x_m](n)]^T$. The actual implementation of (1) and (2) in FXP format gives rise to nonlinear quantization operations that occur at various locations depending on the hardware realization.

Eqn. (1) can be implemented either by using single wordlength accumulators (creating a quantization error after each multiplication) or by using double wordlength accumulators (creating a quantization error only after summation). We will only consider the latter option since practically all modern DSP machines implement this. Eqn. (1) can then be written as

$$\delta[\mathbf{x}](n) = Q\{A^\delta \mathbf{x}(n)\}, \quad (4)$$

where Q is a vector-valued quantization nonlinearity of the form

$$Q\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}. \quad (5)$$

Here, $Q\{x_\nu\}$ denotes magnitude truncation, two's complement truncation, or rounding.

Eqn. (2) can be implemented in two different ways:

$$\mathbf{x}(n+1) = \mathbf{x}(n) + Q\{\Delta \cdot \delta[\mathbf{x}](n)\}, \quad (6)$$

or

$$\mathbf{x} = Q\{\mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n)\}. \quad (7)$$

Eqn. (6) corresponds to quantization after multiplication while (7) corresponds to quantization after summation. In contrast to (1), for (2), it is not clear which of the two quantization schemes in (6) and (7) is preferable. We will therefore consider both possibilities.

Throughout this paper, we will use the following definition of stability:

Definition. The discrete-time system in $\{(4), (6)\}$ or $\{(4), (7)\}$ is globally asymptotically stable if and only if, for any initial condition $\mathbf{x}(0)$, the state vector $\mathbf{x}$ asymptotically reaches zero, that is, $\mathbf{x}(n) \to 0$ for $n \to \infty$.

Comment. Since the FXP systems considered are in fact finite state machines, the condition $\mathbf{x}(n) \to \mathbf{0}$ for $n \to \infty$ may be restated as $\mathbf{x}(N) = \mathbf{0}$ for some finite N [5].

Finally, the symbol ℓ is used to denote the quantization step.

III. NECESSARY CONDITIONS FOR STABILITY

First, we will consider the system described by $\{(4), (6)\}$. From the definition for global asymptotic stability as stated in the previous section, it is necessary that

$$Q\{\Delta \cdot \delta[\mathbf{x}](n)\} \neq \mathbf{0}, \quad \text{for any} \quad \mathbf{x}(n) \neq \mathbf{0}. \quad (8)$$

This is just one of a finite set of conditions that is required to ensure global asymptotic stability of a FXP implementation of a linearly stable system [5].

In the case of rounding, condition (8) is violated if

$$|\Delta \cdot \delta[x_\nu](n)| \leq \frac{\ell}{2}, \quad \text{for any} \quad \nu = 1, \ldots, m. \quad (9)$$

The sampling time Δ in a δ-operator formulated implementation is typically very small. With $\Delta = I \cdot \ell$ and (9), we have

$$|\delta[x_\nu](n)| \leq \frac{1}{2I}, \quad \text{for any} \quad \nu = 1, \ldots, m, \quad (10)$$

where I is a positive integer.

In the case of magnitude truncation, (10) takes the form

$$|\delta[x_\nu](n)| \leq \frac{1}{I}, \quad \text{for any} \quad \nu = 1, \ldots m. \quad (11)$$

Accordingly, for two's complement truncation, we have

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \quad \text{for any} \quad \nu = 1, \ldots, m. \quad (12)$$

Conditions (10-12) describe the deadband, in terms of $\delta[\mathbf{x}]$, for which a DC limit cycle occurs. Such a limit cycle can be avoided if (10-12) are satisfied by the zero vector only. In the case of rounding, we therefore require

$$\ell > \frac{1}{2I},$$

or, equivalently,

$$\Delta > \frac{1}{2}, \quad (13)$$

which is impractical. Similarly, for magnitude and two's complement truncation, we obtain

$$\ell > \frac{1}{I} \Longleftrightarrow \Delta > 1, \quad (14)$$

which again is equally impractical.

This result is summarized in the following theorem.

Theorem 1. A necessary condition for stability of the δ-operator formulated discrete-time system in $\{(4),(6)\}$ is $\Delta > 0.5$ for rounding and $\Delta > 1$ for truncation.

The above theorem shows that high-speed δ-operator formulated implementations that possess a small sampling time cannot be realized limit cycle free in FXP format!

A second necessary condition for the system in $\{(4),(6)\}$ can be obtained by noting that

$$\delta[\mathbf{x}](n) = \mathbf{0} \tag{15}$$

can occur in (4) even though the state vector $\mathbf{x}(n) \neq \mathbf{0}$.

Therefore, for rounding, no nonzero state vector $\mathbf{x}(n)$ that satisfies

$$-\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \leq A^{\delta} \cdot \mathbf{x}(n) \leq +\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \tag{16}$$

may be allowed to exist. Here, the inequality has to hold elementwise. Taking norms on both sides of (16) one gets an algebraic condition on the system matrix A^{δ} that always support DC limit cycles. Eqn. (16) has the following interesting interpretations:

1. Each of the resulting m inequalities can be geometrically interpreted as the intersection of two half spaces in $\Re^m$. These intersections are symmetric about the origin and have parallel boundaries. The normal vector to the boundaries is given by the particular row vector of A^{δ}. Only if the intersection of *all* such m half spaces contains a nonzero point in $\Re^m$, and if it belongs to the quantization lattice, will there exist a nonzero state vector that is an equilibrium point of the system.

2. Eqn. (16) can also be interpreted from an eigenvalue/eigenvector viewpoint. In high-speed digital filters where the sampling frequency is typically much higher than the bandwidth of the processed signal, a q-operator implementation's eigenvalues cluster around the point $z = 1$ [1]. The corresponding δ-operator implementation for large sampling times has eigenvalues clustered around zero. However, as the sampling time becomes small, these eigenvalues move towards the eigenvalues of the underlying continuous-time system [1]. In other words, for large sampling times, the system matrix will be ill-conditioned, that is, vectors $\mathbf{x}(n) \neq \mathbf{0}$ exist such that $A^{\delta} \cdot \mathbf{x}(n)$ is close to the zero vector. According to (16), this is likely to cause a DC limit cycle. For small sampling times, this problem may not occur; however, in this case, the conditions in Theorem 1 are not satisfied!

In the case of the remaining two quantization schemes, the inequalities corresponding to (16) are given as follows: For two's complement truncation,

$$\mathbf{0} \leq A^{\delta} \cdot \mathbf{x}(n) < \begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}, \tag{17}$$

and, for magnitude truncation,

$$-\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix} < A^{\delta} \cdot \mathbf{x}(n) < +\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}. \tag{18}$$

A similar analysis can be conducted for the system in $\{(4),(7)\}$. Since (4) is common to both realizations, (16-18) are still valid and provide conditions under which the finite difference is quantized to zero and a DC limit cycle is produced. We will now briefly discuss necessary conditions for global asymptotic stability obtained from (7).

For rounding, proceeding as in (9), we have

$$|\Delta \cdot \delta[x_\nu](n)| \leq \frac{\ell}{2}, \quad \text{for any} \quad \nu = 1, \ldots, m,$$

and therefore

$$|\delta[x_\nu](n)| \leq \frac{1}{2I}, \quad \text{for any} \quad \nu = 1, \ldots, m. \tag{19}$$

For magnitude truncation, we obtain

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \ \forall \delta[x_\nu] \geq 0, \tag{20}$$

and

$$-\frac{1}{I} < \delta[x_\nu](n) \leq 0, \ \forall \delta[x_\nu] < 0. \tag{21}$$

In the case of two's complement truncation, the condition for a DC limit cycle is given by

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \ \forall \nu = 1, \ldots, m. \tag{22}$$

With $\Delta = I \cdot \ell$, I being a 'small' integer, we come to the same conclusion as for the previously considered system:

$$\Delta > \frac{1}{2} \quad \text{for rounding;}$$

$$\Delta > 1 \quad \text{for truncation.}$$

Therefore, Theorem 1 also holds for the system representation in $\{(4),(7)\}$.

IV. CONCLUSION

Via a set of necessary conditions for global asymptotic stability, it has been shown that high-speed, limit cycle free δ-operator implementations of linear discrete-time systems cannot be realized. This is due to the tendency of such a realization to produce period one limit cycles. This situation arises from small values in the finite difference being quantized to zero. Hence, convergence to the 'wrong' equilibrium point is very likely. Conditions on the system matrix and the sampling time if such limit cycle behavior is to be avoided have been provided. The results indicate that, in high-speed applications, these conditions cannot be satisfied with conventional quantization schemes.

ACKNOWLEDGEMENT

This work was partially supported by a research grant from the Office of Naval Research (ONR).

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High speed digital signal processing and control," *Proceedings of the IEEE,* 80, 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta-operators," *IEEE Transactions Automatic Control,* 31, 11, pp. 1015-1021, Nov. 1986.

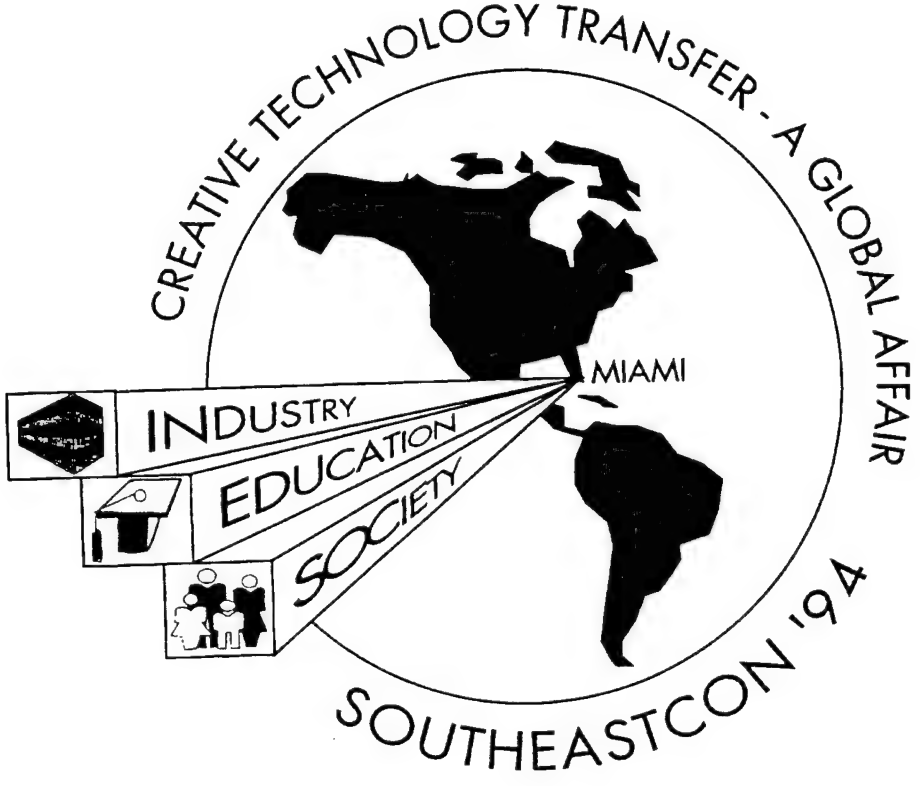
[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the 1990 IEEE Conference on Decision and Control (CDC'90),* 2, pp. 954-959, Honolulu, HI, Dec. 1990.

[4] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Transactions on Signal Processing,* 41, 2, pp. 629-637, Feb. 1993.

[5] P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed point digital filters," *IEEE Transactions on Signal Processing,* 39, 11, pp. 2400-2410, Nov. 1991.

[6] K. Premaratne, R. Salvi, N.R. Habib, and J.P. LeGall, "Delta-operator formulated discrete-time approximations of continuous-time systems," to appear in *IEEE Transactions on Automatic Control,* 1994.

PROCEEDINGS OF 1994 IEEE SOUTHEASTCON '94



Conference and Exhibit

April 10 - 13, 1994
Miami, Florida

Hosted by:
Florida International University ECE Department
University of Miami ECE Department
IEEE Miami Section
IEEE Region 3
IEEE Florida Council

94CH3392-8

FIXED-POINT IMPLEMENTATION OF MULTI-DIMENSIONAL DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS: DIFFICULTIES IN CONVERGENCE

Peter H. Bauer, PhD
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556

Kamal Premaratne, PhD
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124

Abstract— In this paper, the convergence properties of linearly stable multi-dimensional systems are investigated for the case of delta-operator implementations in fixed-point format. It is shown that zero-convergence is almost never achieved, if the sampling time is small. Using a one-dimensional analysis, it is demonstrated that zero-convergence cannot be guaranteed along the axis of the first hyper-quadrant for a first hyper-quadrant causal system. This limits the use of delta-operators for solving partial differential equations in discrete time with fixed-point arithmetic.

I. INTRODUCTION

Delta-operator (or, δ-operator) implementations of discrete-time systems have been the topic of a number of research papers within the last decade. A comprehensive treatment of the properties of δ-operator implementations can be found in [1]. It is well known that δ-operators outperform shift-operators (or, q-operators) in terms of their finite wordlength properties [2]. In particular, its quantization noise and sensitivity properties make the δ-operator an interesting alternative to the q-operator in areas such as digital control, digital signal processing, and generally discrete-time simulation of dynamical systems described by differential equations [1], [3].

In this paper, we will perform a deterministic analysis of the finite wordlength properties of multi-dimensional (m-D) δ-operator implemented discrete-time systems. In particular, we will investigate the zero-convergence of δ-operator fixed-point implementations of one-dimensional (1-D) and m-D systems. Although it is of vital importance, this problem has not been investigated thus far in the literature. After all, asymptotic stability and convergence to the true equilibrium points are some of the most fundamental requirements for any discrete-time system realization.

This article is organized in the following way: Section II introduces the notation. The m-D δ-operator model will be introduced and briefly discussed. This section will also provide the problem formulation. Section III provides necessary 1-D stability conditions for m-D first hyper-quadrant causal systems with nonlinearities. Using these necessary conditions, section IV provides a stability and convergence analysis for m-D systems. It will be shown that the resulting 1-D systems cannot ensure zero-convergence. Section V contains concluding remarks.

II. NOTATION AND PROBLEM FORMULATION

The m-D Roesser model has the following δ-operator formulation [4]:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} A^{\delta}_{11} & \cdots & A^{\delta}_{1m} \\ \vdots & \ddots & \vdots \\ A^{\delta}_{m1} & \cdots & A^{\delta}_{mm} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \begin{bmatrix} B^{\delta}_{1} \\ \vdots \\ B^{\delta}_{m} \end{bmatrix} \mathbf{u}(\mathbf{n}); \quad (1)$$

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix}. \quad (2)$$

The input-state equations in (1) and (2) describe a first hyper-quadrant causal m-D system with a uniform sampling period of Δ in all directions. The operators $q^{(i)}$ and $\delta^{(i)}$ represent the shift- and delta-operator in the direction specified by the axis n_i. In particular

$$q^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) \quad (3a)$$

$$\delta^{(i)}[\mathbf{x}^{(i)}](\mathbf{n})$$

0-7803-1797-1/94/$3.00 © 1994 IEEE

$$= \frac{1}{\Delta}(\mathbf{x}^{(i)}(n_1,\ldots,n_{i-1},n_i+1,n_{i+1},\ldots,n_m) - \mathbf{x}^{(i)}(\mathbf{n})). \quad (3b)$$

Here, $(\mathbf{n}) \doteq (n_1,\ldots,n_m)$ denotes a point in the first hyper-quandrant, $\mathbf{x}^{(i)}(\mathbf{n})$ is the portion of the state vector propagating in the direction specified by the axis n_i, $u(\mathbf{n})$ is the m-D input vector, and A_{ij}^{δ} and B_i^{δ}, for $i = 1,\ldots,m$, $j = 1,\ldots,m$, are the submatrices of the system and input matrices, respectively.

If (1) is realized in fixed-point arithmetic, it takes the following form under zero-input conditions:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} \right\} \quad (4)$$

where $\mathbf{Q}\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}$ with $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$.

Equation (4) assumes quantization after summation; since practically all modern DSP machines implement this quantization scheme, we utilize this. The vector-valued quantization nonlinearity $\mathbf{Q}\{\cdot\}$ may represent any one of the conventional schemes, viz., magnitude truncation, magnitude rounding, two's complement truncation, and two's complement rounding.

Equation (2) can be implemented in two different forms:

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(i)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \quad (5)$$

or

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \quad (6)$$

Equation (5) corresponds to quantization after multiplication, whereas (6) corresponds to quantization after addition. In contrast to (1), for (2), it is not obvious which of the two forms stated above is preferable.

The following definition for asymptotic stability [5] will be used throughout this paper.

Definition. An m-D first hyper-quadrant causal discrete-time system is asymptotically stable under all finitely extended bounded input signals $u(\mathbf{n})$ where

$$|u(\mathbf{n})| \le S, \quad \text{for} \quad n_1 + \cdots + n_m \le D; \quad (7)$$

$$u(\mathbf{n}) = 0, \quad \text{for} \quad n_1 + \cdots + n_m > D, \quad (8)$$

if all the states of the m-D discrete-time system asymptotically reach zero for $n_1 + \cdots + n_m \to \infty$. Here, $n_\nu \ge 0$, $\nu = 1,\ldots,m$, S is a nonnegative real number, and D is a positive integer.

Since the fixed-point systems considered are in fact finite state machines, the condition

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} \to \mathbf{0},$$

for $n_1 + \cdots + n_m \to \infty$, $n_\nu \ge 0$, $\nu = 1,\ldots,m$, can be strengthened to

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} = \mathbf{0},$$

for all points $n_1 + \cdots + n_m \ge c$, $n_\nu \ge 0$, $\nu = 1,\ldots,m$, where c is some finite integer.

Problem Formulation. Analyze the asymptotic zero-convergence of the state response of systems in (4,5) and (4,6) under the assumption that the underlying linear system is asymptotically stable.

III. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY OF m-D SYSTEMS

In this section, we present some necessary conditions for stability of a first hyper-quadrant causal m-D discrete-time system represented in its Roesser local state-space model in (1,2). These necessary conditions are formulated in terms of 1-D conditions. This theorem follows directly from a result in [6] which was formulated for q-operator implemented discrete-time systems. The proof of the theorem rests on the fact that a first hyper-quadrant m-D system can be described by a 1-D system for those locations that are along the m coordinate axes of the boundary of the hyper-quadrant. Reformulating the result in [6] for δ-operator systems produces the following theorem:

Theorem 1.

(a) A necessary condition for global asymptotic stability of the system in (4,5) is that each of the following 1-D systems in (9,10) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{9}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{x}^{(i)}(n_i) + \mathbf{Q}\left\{\Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\} \tag{10}$$

where $i = 1, \ldots, m$.

(b) A necessary condition for global asymptotic stability of the system in (4,6) is that each of the following in 1-D systems in (11,12) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{11}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{\mathbf{x}^{(i)}(n_i) + \Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\} \tag{12}$$

where $i = 1, \ldots, m$.

Proof. For a detailed proof, and generalizations to higher sub-dimensional systems, the reader is referred to [6]. ■

Theorem 1 can be viewed as an extension of the concept of practical BIBO stability to asymptotic stability of nonlinear systems. It is particularly useful in proving instability in m-D nonlinear systems.

IV. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY OF 1-D SYSTEMS

Let us rewrite (9), (10), and (12) as 1-D matrix equations of order K. In this case, (9), (10), and (12) yield (13), (14), and (15), respectively:

$$\begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} a_{11}^{\delta} & \cdots & a_{1K}^{\delta} \\ \vdots & \ddots & \vdots \\ a_{n1}^{\delta} & \cdots & a_{KK}^{\delta} \end{bmatrix} \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} \right\}; \tag{13}$$

$$\begin{bmatrix} x_1(n+1) \\ \vdots \\ x_K(n+1) \end{bmatrix} = \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} \right\}; \tag{14}$$

$$\begin{bmatrix} x_1(n+1) \\ \vdots \\ x_K(n+1) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} \right\}. \tag{15}$$

Now, we are in a position to formulate the second theorem which presents a necessary condition for stability of 1-D systems.

Theorem 2. A necessary condition for global asymptotic stability of the system in (13,14) or (13,15) is given by

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$
$$\Delta \geq 1, \quad \text{for truncating.}$$

Proof. For global asymptotic stability of (13,14), it is necessary that

$$\mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} \right\} \neq 0, \tag{16}$$

for any $\begin{pmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{pmatrix} \neq 0$.

First, we will address the case of magnitude rounding. Obviously, condition (16) is violated if, for $x_\nu \neq 0$,

$$|\Delta \cdot \delta[x_\nu](n)| < \frac{\ell}{2}, \quad \text{for} \quad \nu = 1, \ldots, K, \tag{17}$$

where ℓ is the quantization step. Expressing the sampling time Δ as an integer multiple of ℓ, we have

$$\Delta = I \cdot \ell, \tag{18}$$

where I is some (typically small) positive integer. With (17) and (18), we obtain the following condition for instability:

$$|\delta[x_\nu](n)| < \frac{1}{2I}, \ \nu = 1, \ldots, m, \tag{19}$$

for $x_\nu \neq 0$, $\nu = 1, \ldots, m$.

Condition (19) is not satisfied for any nonzero value of x_ν (that is, the condition for instability is not satisfied) if $\ell \geq 1/2I$, or equivalently,

$$\Delta \geq \frac{1}{2}. \tag{20}$$

This proves the theorem for magnitude rounding.

For the case of magnitude truncating, (17) takes the form

$$|\Delta \cdot \delta[x_\nu](n)| < \ell, \quad \text{for} \quad \nu = 1, \ldots, K. \tag{21}$$

Therefore, (19) becomes

$$|\delta[x_\nu](n)| < \frac{1}{I}. \tag{22}$$

This finally yields

$$\Delta \geq 1. \tag{23}$$

For two's complement, (17) takes the form

$$0 \leq \Delta \cdot \delta[x_\nu](n) < \ell, \quad \text{for} \quad \nu = 1, \ldots, K. \tag{24}$$

This results in

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \tag{25}$$

and consequently, $\Delta \geq 1$. This proves the theorem for the system in (13,14). A similar argument can be used for the system in (13,15) by considering the cases for which

$$\mathbf{Q}\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix}\right\} = \mathbf{Q}\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix}\right\}, \tag{26}$$

for nonzero state vectors. ■

We can now combine Theorems 1 and 2 to formulate a necessary condition for stability of m-D first hyperquadrant causal δ-operator formulations of the generalized Roesser model.

Corollary 3. A necessary condition for global asymptotic stability of the m-D systems in (4,5) or (4,6) is

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1, \quad \text{for truncatiing.}$$

Proof. The proof follows from Theorems 1 and 2. ■

Comments.

1. Theorem 2 and Corollary 3 are also essentially applicable to the case where the sampling time varies with the direction of propagation. In this case, the inequalities in Theorem 2 and Corollary 3 would have to be replaced by

   $$\Delta_i \geq 0.5, \quad \text{for magnitude rounding;}$$

   $$\Delta_i \geq 1, \quad \text{for truncating,}$$

   for $i = 1, \ldots, m$.
2. Most of the previous results on the superior finite wordlength properties of δ-operators depend on choosing a very small sampling time Δ. In such a case, Theorem 2 and Corollary 3 show that the system response will not converge to zero for the unforced case.
3. Our analysis is limited to the zero-input case for which DC limit cycles were used to derive conditions for non-convergence. If one includes other types of limit cycles in the analysis, the requirements for Δ may become even more severe.
4. Theorem 2 and Corollary 3 show that fixed-point implementations of 1-D and m-D δ-operator systems *cannot be realized limit cycle free, if good coefficient sensitivity and quantization noise measures have to be achieved.* See also [7].

V. CONCLUSION

In this paper, it was shown that fixed-point implementations of 1-D and m-D δ-operator systems are not limit cycle free even if the underlying linear system is stable and the sampling time is chosen small. This nonconvergent behavior can be explained by the quantization of the δ-term to zero which leaves the state vector unchanged. The smaller the sampling time, the more severe this effect is. Therefore, the practical value of δ-operators for fixed-point implementations of 1-D and m-D systems is questionable. There are however indications that this effect is much less severe in floating-point implementations.

δ-operator implemented discrete-time systems represent a class of systems where the quantization noise at the output can be small compared to other realizations. However, as was shown above, such realizations will invariably exhibit limit cycle, that is, highly correlated quantization noise, behavior. Therefore, in this case, typical measures for quantization noise are of very limited use for obtaining any insight into the likelihood of limit cycles and vice versa.

ACKNOWLEDGEMENT

This work was partially supported by a grant from the Office of Naval Research (ONR).

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE*, vol. 80, no. 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta operators," *IEEE Transactions on Automatic Control*, vol. 31, pp. 1015-1021, Nov. 1986.

[3] G.Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the IEEE Conference on Decision and Control (CDC'90)*, vol. 2, pp. 954-959, Honolulu, HI, 1990.

[4] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer, "Delta-operator formulated implementation of two-dimensional discrete-time systems," in preparation.

[5] P. Bauer, "Finite wordlength effects in m-D digital filters with singularities on the stability boundary," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 894-900, Apr. 1992.

[6] P. Bauer, "A set of necessary stability conditions for m-D nonlinear digital filters," to appear in *Circuits, Systems and Signal Processing*, 1994.

[7] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," submitted to be presented at the *1994 IEEE Symposium on Circuits and Systems (ISCAS'94)*, London, UK, 1994.

To appear in Multidimensional Systems and Signal Processing, 1994.

jun0193*.tex

An Algorithm for Stability Determination of Two-Dimensional Delta-Operator Formulated Discrete-Time Systems

KAMAL PREMARATNE
Department of Electrical and Computer Engineering, University of Miami, P.O. Box 248294, Coral Gables, FL 33124, U.S.A.

A.S. BOUJARWAH
Electrical and Computer Engineering Department, College of Engineering and Petroleum, Kuwait University, P.O. Box 5969, 13060 Safat, Kuwait.

Abstract. The recent interest in delta-operator (or, δ-operator) formulated discrete-time systems (or, δ-systems) is due mainly to (a) their superior finite wordlength characteristics as compared to their more conventional shift-operator (or, q-operator) counterparts (or, q-systems), and (b) the possibility of a more unified treatment of both continuous- and discrete-time systems. With such advantages, design, analysis, and implementation of two-dimensional (2-D) discrete-time systems using the δ-operator is indeed warranted. Towards this end, the work in this paper addresses the development of an easily implementable *direct* algorithm for stability checking of 2-D δ-system transfer function models. *Indirect* methods that utilize transformation techniques are not pursued since they can be numerically unreliable. In developing such an algorithm, a tabular form for stability checking of δ-system characteristic polynomials with complex-valued coefficients and certain quantities that may be regarded as their corresponding Schur-Cohn minors are also proposed.

Keywords. Two-dimensional discrete-time systems, two-dimensional digital filters, δ-operator formulated discrete-time systems, bivariate polynomials, Schur-Cohn minors, stability.

1. Introduction

The increased interest in δ-systems during the recent years (see [1-6], and references therein) is due mainly to two reasons: (a) δ-systems provide superior finite wordlength properties with respect to roundoff noise propagation [5] and coefficient sensitivity [1], [5], [7], as compared to their q-system counterparts, and (b) the δ-operator yields the differential operator as a limiting case when sampling time approaches zero enabling a unified treatment of both continuous- and discrete-time systems [1].

With such advantages in mind, development of 2-D and multi-dimensional (m-D) δ-system models must clearly be undertaken. Such research can, for example, provide m-D digital filters with superior roundoff error and coefficient sensitivity performance allowing their implementation to be carried out in a shorter wordlength environment. This is especially crucial in real-time applications, such as, in implementing narrow bandwidth filters under high sampling rates (for example, in current wide bandwidth communication system applications) where traditional q-operator implementations perform poorly [8].

In applications mentioned above, and those dealing with high-speed processing of 2-D and m-D data (for instance, in weather, seismic, gravitational photographs, video images, systems with mutliple sampling rates, etc.), ensuring stability is an important consideration (see [9], and references therein). Given the characteristic polynomial of a δ-system, to determine stability, one may first use a variable transformation that yields a more familiar stability region, for instance, the unit bi-circle. Then, an existing technique (see [9-10], and references therein) may be applied. However, such techniques are known to be prone to numerically ill-conditioning [1], [6]. In the 1-D case, direct stability checking methods for δ-system polynomials are in [6] (where a tabular method based on the work in [11] is given) and [12] (where a Hermite-Bieler-like Theorem is utilized). Hence, our purpose here is to develop a *direct* easily implementable stability checking technique applicable to m-D δ-systems. As usual, for notational simplicity, we concentrate on the 2-D case, the extension to the m-D case being quite straight-forward.

In checking stability of bivariate characteristic polynomials, two conditions must be

satisfied.

(a) Condition I involves a 1-D stability check of a polynomial with real-valued coefficients. One may use the table form in [6]. Alternately, one may utilize an explicit root location scheme.

(b) Condition II involves a stability check of a polynomial with complex-valued coefficients where the latter are dependent on a parameter taking values on a certain circle in the complex plane. Explicit root location schemes are now ineffective, and the value of tabular methods becomes apparent. Note that, in such a situation, compared to Nyquist-like techniques [13], tabular methods are known to provide certain numerical advantages as well [14].

In checking condition II for 2-D q-systems, an effective technique involves checking positive definiteness of the Hermitian Schur-Cohn matrix [15]. This lets one use an important simplification due to Siljak [16]. The tabular form in [15] takes full use of this since it provides the Schur-Cohn minors (that is, the principal minors of the Hermitian Schur-Cohn matrix) directly from its entries [15], [17]. A similar simplification applicable to δ-systems is clearly possible if condition II may be reduced to checking positive definiteness of a Hermitian matrix.

With the above in mind, we develop the following in this paper: (a) Tabular form for stability checking of δ-system characteristic polynomials possessing complex-valued coefficients, (b) Analogs of Schur-Cohn minors and a corresponding Hermitian matrix applicable for such systems, and (c) a direct stability checking algorithm for 2-D δ-system transfer function models.

The paper is organized as follows. Section 2 introduces the notation used throughout and a brief review of previous results. Section 3 develops a tabular form for stability checking of δ-systems with complex-valued coefficients and some important relevant results. Section 4 presents quantities that may be regarded as the analogs of Schur-Cohn minors for δ-systems. The 2-D stability checking algorithm in Section 5 is based on the tabular form for real-valued coefficients [6]. Since only little extra work is needed, results in both

Sections 3 and 4 however are developed for the more general complex-coefficient case. Section 6 presents an example to validate the results. Section 7 contains the conclusion and some final remarks.

2. Preliminaries

2.1. Notation

$\Re$, $\Im$	Real and complex number fields.
$\Re^{p\times q}$, $\Im^{p\times q}$	Set of matrices of size $p\times q$ over $\Re$ and $\Im$, respectively.
$\mathrm{var}\{\cdot\}$	Number of sign changes in the sequence $\{\cdot\}$ of real numbers.
$\mathrm{Re}[\cdot]$, $\mathrm{Im}[\cdot]$	Real part and imaginary part of $[\cdot]\in\Im$.
$\bar{[\cdot]}$	Complex conjugate of $[\cdot]\in\Im$.
A^T, $\bar{A}$, A^*	Transpose, complex conjugate, and complex conjugate transpose of $A\in\Im^{p\times q}$, respectively.
$\Re[w]_n$, $\Im[w]_n$	Set of univariate polynomials of degree n (with respect to the indeterminate $w\in\Im$) over $\Re$ and $\Im$, respectively.
$\Re(w)$	Set of rational univariate polynomials (that is, quotient of univariate polynomials) over $\Re$.
$\Re[w_1]_{n_1}[w_2]_{n_2}$	Set of bivariate polynomials of relative degrees n_1 and n_2 (with respect to the indeterminates $w_1\in\Im$ and $w_2\in\Im$, respectively) over $\Re$.
$\Re(w_1,w_2)$	Set of rational bivariate polynomials over $\Re$.
z, c	Indeterminates of q- and δ-systems, respectively.
τ	Real positive number, usually the sampling time.

The transformation relationship between corresponding q-and δ-systems is

$$\delta=\frac{q-1}{\tau}\Longleftrightarrow c=\frac{z-1}{\tau}. \tag{2.1}$$

$\breve{[\cdot]}$	q-system quantity analogous to its corresponding δ-system quantity $[\cdot]$; for example, transfer function of a given discrete-time system is either $H(c)$ if implemented based on the δ-operator or $\breve{H}(z)$ if implemented based on the q-operator.
$H(c)\vert_{c\to z}$	$H(c)\vert_{c=(z-1)/\tau}$
$G(z)\vert_{z\to c}$	$G(z)\vert_{z=1+\tau c}$
$H(c_1,c_2)\vert_{\mathbf{c}\to\mathbf{z}}$	$H(c_1,c_2)\vert_{c_i=(z_i-1)/\tau,\,i=1,2}$
$G(z_1,z_2)\vert_{\mathbf{z}\to\mathbf{c}}$	$G(z_1,z_2)\vert_{z_i=1+\tau c_i,\,i=1,2}$

Stability studies of 1-D and 2-D q- and δ-systems involve the following regions:

$\mathcal{U}_q, \mathcal{U}_q^2$ $\{z \in \Im : |z| < 1\}$, $\{(z_1, z_2) \in \Im^2 : |z_i| < 1, i = 1, 2\}$.

$\overline{\mathcal{U}}_q, \overline{\mathcal{U}}_q^2$ $\{z \in \Im : |z| \leq 1\}$, $\{(z_1, z_2) \in \Im^2 : |z_i| \leq 1, i = 1, 2\}$.

$\mathcal{T}_q, \mathcal{T}_q^2$ $\{z \in \Im : |z| = 1\}$, $\{(z_1, z_2) \in \Im^2 : |z_i| = 1, i = 1, 2\}$.

$\mathcal{U}_\delta, \mathcal{U}_\delta^2$ $\{c \in \Im : |c + 1/\tau| < 1/\tau\}$, $\{(c_1, c_2) \in \Im^2 : |c_i + 1/\tau| < 1/\tau, i = 1, 2\}$.

$\overline{\mathcal{U}}_\delta, \overline{\mathcal{U}}_\delta^2$ $\{c \in \Im : |c + 1/\tau| \leq 1/\tau\}$, $\{(c_1, c_2) \in \Im^2 : |c_i + 1/\tau| \leq 1/\tau, i = 1, 2\}$.

$\mathcal{T}_\delta, \mathcal{T}_\delta^2$ $\{c \in \Im : |c + 1/\tau| = 1/\tau\}$, $\{(c_1, c_2) \in \Im^2 : |c_i + 1/\tau| = 1/\tau, i = 1, 2\}$.

To avoid unnecessary notational complications, the sampling time in both horizontal and vertical directions is taken to be equal to $\tau > 0$.

To emphasize the degree of $F(w) = \sum_{k=0}^{n} a_k^{(n)} w^k \in \Im[w]_n$, we sometimes denote it as $F(w)_n$ as well.

$\bar{F}(w)$ Conjugate polynomial of $F(w)$, that is, $\sum_{k=0}^{n} \bar{a}_k^{(n)} w^k$

$F^\sharp(z)$ Reciprocal polynomial of $F(z)$, that is, $z^n \bar{F}(1/z)$

$F^\sharp(c)$ Reciprocal polynomial of $F(c)$, that is, $(1 + \tau c)^n \bar{F}\left(\frac{-c}{1+\tau c}\right)$

A q-system polynomial is *q-symmetric* if $F(z) = F^\sharp(z)$. A δ-system polynomial is *δ-symmetric* if $F(c) = F^\sharp(c)$.

Tabular forms of stability checking of a polynomial in $\Im[\omega]_n$ typically employ a sequence of polynomials each of descending order. The first row of such a tabular form is denoted as *row* $\#n$, the second row is *row* $\#n - 1$, and so on.

JT, MJT Jury table [18], modified Jury table [15], [17].

real-q-BT Bistritz table for q-system polynomials with real-valued coefficients [11].

complex-q-BT Bistritz table for q-system polynomials with complex-valued coefficients [19].

real-δ-BT Table form for δ-system polynomials with real-valued coefficients [6].

complex-δ-BT Table form for δ-system polynomials with complex-valued coefficients (this paper).

A q-system polynomial with all its roots in $\mathcal{U}_q$ (for the 1-D case) or $\mathcal{U}_q^2$ (for the 2-D case) is said to be *stable*. The corresponding regions for a δ-system polynomial are $\mathcal{U}_\delta$ (for the 1-D case) or $\mathcal{U}_\delta^2$ (for the 2-D case), respectively.

2.2. Review of complex-q-BT

The complex-δ-BT introduced in Section 3 is based on the complex-q-BT, and hence, we briefly review it now. For more details, see [10]. Let the characteristic polynomial of a q-system be

$$\breve{F}(z) = \sum_{k=0}^{n} \breve{a}_k^{(n)} z^k \in \Im[z]_n \quad \text{with} \quad \breve{F}(1) \in \Re \quad \text{and} \quad \breve{F}(1) \neq 0. \tag{2.2}$$

The complex-q-BT is formed using the symmetric polynomial sequence $\{\breve{T}(z)_i\}_{i=0}^n$ where [19]

$$\breve{T}(z)_i = \begin{cases} \breve{F}(z)_n + \breve{F}^\sharp(z)_n, & \text{for } i = n; \\ \dfrac{\breve{F}(z)_n - \breve{F}^\sharp(z)_n}{z-1}, & \text{for } i = n-1; \\ \dfrac{(\breve{\delta}_{i+2} + \bar{\breve{\delta}}_{i+2} z)T(z)_{i+1} - T(z)_{i+2}}{z}, & \text{for } i = n-2, n-3, \ldots, 0, \end{cases} \tag{2.3}$$

where

$$\breve{\delta}_{i+2} = \frac{\breve{T}(0)_{i+2}}{\breve{T}(0)_{i+1}} = \frac{\breve{t}_0^{(i+2)}}{\breve{t}_0^{(i+1)}}, \; i = n-2, n-3, \ldots, 0. \tag{2.4}$$

As in [11] and [19], equating similar powers on either side, we may also get the following determinental rule: For $k = 0, 1, \ldots, i$, and $i = n-2, n-3, \ldots, 0$,

$$\breve{t}_k^{(i)} = \frac{1}{\breve{t}_0^{(i+1)}} \begin{vmatrix} \breve{t}_0^{(i+2)} & \breve{t}_{k+1}^{(i+2)} \\ \breve{t}_0^{(i+1)} & \breve{t}_{k+1}^{(i+1)} \end{vmatrix} + \frac{1}{\breve{t}_{i+1}^{(i+1)}} \begin{vmatrix} \breve{t}_{i+2}^{(i+2)} & \breve{t}_{k+1}^{(i+2)} \\ \breve{t}_{i+1}^{(i+1)} & \breve{t}_k^{(i+1)} \end{vmatrix} + \breve{t}_{k+1}^{(i+2)}. \tag{2.5}$$

Remark. The computational advantage of BT is due to $\breve{T}(z)_i$ being q-symmetric. This implies $\breve{t}_k^{(i)} = \bar{\breve{t}}_{i-k}^{(i)}$, $k = 0, 1, \ldots, i$, and hence, it is necessary to evaluate only half the coefficients of each row.

Using (12-13), (16), and Theorem 6 of [19], we get

THEOREM 2.1. [19] The polynomial $\breve{F}(z) \in \Im[z]_n$ is q-stable iff
I. $\breve{t}_0^{(i)} \neq 0$, $i = n-1, n-2, \ldots, 0$, and
II. $\nu_n = \text{var}\{\breve{T}(1)_n, \breve{T}(1)_{n-1}, \ldots, \breve{T}(1)_0\} = 0$.

2.3. Some results on 2-D stability

Consider the 2-D q-system transfer function

$$\breve{H}(z_1, z_2) = \frac{\breve{E}(z_1, z_2)}{\breve{F}(z_1, z_2)} \in \Re(z_1, z_2) \tag{2.6}$$

where $\breve{E}(z_1, z_2) \in \Re[z_1]_{n_1}[z_2]_{n_2}$ and $\breve{F}(z_1, z_2) \in \Re[z_1]_{n_1}[z_2]_{n_2}$. The 2-D z-transform is taken using positive powers of z_i. For a comprehensive discussion regarding stability of such systems, see [9-10], and references therein. Hence, for reasons of brevity, only some analog results applicable to 2-D δ-systems are provided. It is only necessary to observe that the corresponding δ-system $H(c_1, c_2)$ satisfies

$$H(c_1, c_2) = \frac{E(c_1, c_2)}{F(c_1, c_2)} = \breve{H}(z_1, z_2)|_{\mathbf{z}\to\mathbf{c}} \in \Re(c_1, c_2) \tag{2.7}$$

where $E(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$ and $F(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$. In the remainder of this paper, we will only be dealing with transfer functions $H(c_1, c_2)$ that are devoid of nonessential singularities of the second kind on $\mathcal{T}_\delta^2$ and the pair $E(c_1, c_2)$ and $F(c_1, c_2)$ is taken to be coprime. If the 2-D polynomial $F(c_1, c_2) \neq 0$, $\forall (c_1, c_2) \in \overline{\mathcal{U}}_\delta^2$, it is said to be *$\delta$-stable*. After using (2.1), the following result follows directly from [20]:

THEOREM 2.2. The 2-D δ-system in (2.7) is δ-stable iff
I. $F(c_1, -1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and
II. $F(c_1, c_2) \neq 0$, $\forall c_1 \in \mathcal{T}_\delta$, $\forall c_2 \in \overline{\mathcal{U}}_\delta$.

The following result, which allows one to use the real-δ-BT, is directly from [21-22] after using (2.1):

THEOREM 2.3. The 2-D δ-system in (2.7) is δ-stable iff
I. $F(c_1, -1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and

II. $G(x, c_2) \neq 0$, $\forall x \in [-2/\tau, 0]$, $\forall c_2 \in \overline{\mathcal{U}}_\delta$.
Here $G(x, c_2) = F(c_1, c_2)F(\bar{c}_1, c_2)\Big|_{\substack{c_1 \in \mathcal{T}_\delta \\ x=(c_1+\bar{c}_1)/2}}$.

2.4. Schur-Cohn minors

In stability checking of 2-D q-systems, the following result is important:

THEOREM 2.4. [15], [23-24] The polynomial $\breve{F}(z) \in \Im[z]_n$ is stable iff $\breve{\Delta}_i > 0$, $i = 1, 2, \ldots, n$, where $\breve{\Delta}_i$ is the principal minor of the Hermitian Schur-Cohn matrix $\breve{\Gamma} = \breve{\Gamma}^* = \{\breve{\gamma}_{ij}\} \in \Im^{n\times n}$ defined as

$$\breve{\gamma}_{ij} = \sum_{k=1}^{i} (\breve{a}_{n-i+k}\bar{\breve{a}}_{n-j+k} - \bar{\breve{a}}_{i-k}\breve{a}_{j-k}), \quad \text{for} \quad i \leq j.$$

Stability checking of 2-D q-systems then involves positivity checking of all Schur-Cohn minors $\breve{\Delta}_i(z)$, $\forall i = 1, 2, \ldots, n$, $\forall |z| = 1$. A necessary and sufficient condition for this is positivity of $\breve{\Delta}_i(1)$, $\forall i = 1, 2, \ldots, n$, and $\breve{\Delta}_n(z)$, $\forall |z| = 1$. This is the simplification due to [16] that has been effectively utilized in applying the MJT [15]. The advantage of the latter is that its entries yield the Schur-Cohn minors directly. The fact that complex-q-BT's entries also yield the Schur-Cohn minors was only recently shown.

THEOREM 2.5. [10], [25] The Schur-Cohn minors of $\breve{F}(z)$ are the principal minors of the $(n \times n)$ tridiagonal Hermitian matrix

$$\breve{\Delta} = \begin{bmatrix} \mathrm{Re}[\breve{t}_0^{(n)}\breve{t}_{n-1}^{(n-1)}] & \frac{1}{2}[\breve{t}_{n-1}^{(n-1)}\breve{t}_0^{(n-2)}] & 0 & \cdots & 0 & 0 \\ \frac{1}{2}[\breve{t}_0^{(n-1)}\breve{t}_{n-2}^{(n-2)}] & \mathrm{Re}[\breve{t}_0^{(n-1)}\breve{t}_{n-2}^{(n-2)}] & \frac{1}{2}[\breve{t}_{n-2}^{(n-2)}\breve{t}_0^{(n-3)}] & \ddots & 0 & 0 \\ 0 & \frac{1}{2}[\breve{t}_0^{(n-2)}\breve{t}_{n-3}^{(n-3)}] & \mathrm{Re}[\breve{t}_0^{(n-2)}\breve{t}_{n-3}^{(n-3)}] & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \mathrm{Re}[\breve{t}_0^{(2)}\breve{t}_1^{(1)}] & \frac{1}{2}[\breve{t}_1^{(1)}\breve{t}_0^{(0)}] \\ 0 & 0 & 0 & \cdots & \frac{1}{2}[\breve{t}_0^{(1)}\breve{t}_0^{(0)}] & \mathrm{Re}[\breve{t}_0^{(1)}\breve{t}_0^{(0)}] \end{bmatrix}.$$

3. Complex-δ-BT

With no loss of generality, consider the δ-system characteristic polynomial

$$F(c) = \sum_{k=0}^{n} a_k^{(n)} c^k \in \Im[c]_n, \tag{3.1}$$

where

$$a_0^{(n)} \in \Re \quad \text{and} \quad a_0^{(n)} > 0. \tag{3.2}$$

We now construct the complex-δ-BT with the use of the δ-symmetric polynomial sequence $\{T(c)_i\}_{i=0}^{n}$ where

$$T(c)_i = \begin{cases} F(c)_n + F^{\sharp}(c)_n, \ i = n; \\ \dfrac{F(c)_n - F^{\sharp}(c)_n}{c}, \ i = n-1; \\ \dfrac{(\delta_{i+2} + \bar{\delta}_{i+2}(1+\tau c))T(c)_{i+1} - T(c)_{i+2}}{1+\tau c}, \ i \leq n-2. \end{cases} \tag{3.3}$$

Here

$$\delta_{i+2} = \frac{T(-1/\tau)_{i+2}}{T(-1/\tau)_{i+1}}, \ i = n-2, n-3, \ldots, 0. \tag{3.4}$$

The *normal conditions* required to complete the sequence are

$$T(-1/\tau)_i \neq 0, \ i = 1, 2, \ldots, n-1. \tag{3.5}$$

Remarks.

1. To determine δ-stability of $F(c)$, one may of course first obtain $\breve{F}(z) = F(c)|_{c \to z}$ and then determine its q-stability by applying familiar stability checking algorithms (e.g., BT or MJT). The possible shortcomings of such a scheme are outlined in [1] and [6]. The purpose here is to obtain a direct check for δ-stability.

2. We follow the work in [6] and [19], and hence, for brevity, all details are omitted.

3. The conditions $T(-1/\tau)_i = 0$, for some $i = 1, 2, \ldots n-1$, imply certain singular conditions on the root distribution of $F(c)$ [11], [19]. The equivalent singular conditions for the real-δ-BT is in [6].

4. Using δ-symmetry, it is easy to show that

$$T(-1/\tau)_i = \frac{\bar{t}_i^{(i)}}{\tau^i}, \ i = 0, 1, \ldots, n. \tag{3.6}$$

Therefore

$$\delta_{i+2} = \frac{1}{\tau} \frac{t_{i+2}^{(i+2)}}{t_{i+1}^{(i+1)}}, \; i = n-2, n-3, \ldots, 0. \tag{3.7}$$

The normal conditions in (3.5) may now be expressed as

$$t_{i+1}^{(i+1)} \neq 0, \; i = n-2, n-3, \ldots, 0. \tag{3.8}$$

Analogous to [6], [11], and [19], we then have

THEOREM 3.1. The polynomial $F(c) \in \Im[c]_n$ is stable iff
I. $t_i^{(i)} \neq 0, \; i = n-1, n-2, \ldots, 1$, and
II. $\nu_n = \text{var}\{T(0)_n, T(0)_{n-1}, \ldots, T(0)_0\} = 0$.

One of the main advantages of the complex-q-BT is that all computations may be carried out through real arithmetic only [19]. The same holds true for the the complex-δ-BT introduced above as well. To see this, let

$$T(c)_i = S(c)_i + jA(c)_i \quad \text{with} \quad \delta_i = \text{Re}[\delta_i] + j\text{Im}[\delta_i], \tag{3.9}$$

for $i = 2, 3, \ldots, n$. It is easy to show that $S(c)_i$'s and $A(c)_i$'s form sequences of δ-symmetric and δ-antisymmetric polynomials, respectively. Now, (3.3) may be expressed as

$$\begin{aligned} S(c)_{i-2} &= \frac{1}{1+\tau c} \left[\text{Re}[\delta_i](2+\tau c) \cdot S(c)_{i-1} + \text{Im}[\delta_i]\tau c \cdot A(c)_{i-1} - S(c)_i\right]; \\ A(c)_{i-2} &= \frac{1}{1+\tau c} \left[-\text{Im}[\delta_i]\tau c \cdot S(c)_{i-1} + (2+\tau c)\text{Re}[\delta_i] \cdot A(c)_{i-1} - A(c)_i\right], \end{aligned} \tag{3.10}$$

for $i = 2, 3, \ldots, n$.

Remark. Note that, $T(0)_i = S(0)_i + jA(0)_i = S(0)_i$.

In the real-δ-BT construction, a certain 'scaling' of $\{T(c)_i\}_{i=0}^n$ was useful [6]. We use the same technique in the complex-δ-BT case as well, thus providing the following advantages: (a) Terms containing τ are avoided during construction, (b) δ_i and ν_i may be deduced by simple inspection, and thus (c) computational effort is reduced.

The sequence of polynomials that incorporates 'scaling' is $\{U(\zeta)_i\}_{i=0}^n$ where

$$U(\zeta)_i = \sum_{k=0}^{i} u_k^{(i)} \zeta^k = T(c)_i \Big|_{c=-\zeta/\tau} \Longleftrightarrow u_k^{(i)} = \left(-\frac{1}{\tau}\right)^k t_k^{(i)},\ k = 0, 1, \ldots, i, \tag{3.11}$$

for $i = 0, 1, \ldots, n$. Thus, from (3.3), we get, for $i = n-2, n-3, \ldots 0$,

$$\begin{aligned} u_0^{(i)} &= (\delta_{i+2} + \bar{\delta}_{i+2}) u_0^{(i+1)} - u_0^{(i+2)}; \\ u_k^{(i)} &= (\delta_{i+2} + \bar{\delta}_{i+2}) u_k^{(i+1)} - \bar{\delta}_{i+2} u_{k-1}^{(i+1)} - u_k^{(i+2)} + u_{k-1}^{(i)},\ k = 1, 2, \ldots i. \end{aligned} \tag{3.12}$$

Note that

$$\delta_{i+2} = \frac{1}{\tau} \frac{\bar{t}_{i+2}^{(i+2)}}{\bar{t}_{i+1}^{(i+1)}} = -\frac{\bar{u}_{i+2}^{(i+2)}}{\bar{u}_{i+1}^{(i+1)}},\ i = n-2, n-3, \ldots, 0, \tag{3.13}$$

and

$$\nu_n = \mathrm{var}\{T(0)_i\}_{i=0}^n = \mathrm{var}\{u_0^{(i)}\}_{i=0}^n. \tag{3.14}$$

Therefore, condition II of Theorem 3.1 may be checked by inspecting the constant coefficients of $\{U(\zeta)_i\}_{i=0}^n$.

Remark. One may use the same 'scaling' strategy in an implementation that uses only real arithmetic.

Relationship between complex-q-BT and complex-δ-BT

As was agreed upon previously, given $F(c)_n \in \Im[z]$, let us use the notation $\breve{F}(z)_n$ to indicate

$$\breve{F}(z)_n = \lambda F(c)_n \Big|_{c \to z} \tag{3.15}$$

where $\lambda \in \Re$ is a possible scaling constant. The establishment of the relationship between the rows of complex-q-BT of $\breve{F}(z)$, i.e., $\{\breve{T}(z)_i\}_{i=0}^n$, and complex-$\delta$-BT of $F(c)$, i.e., $\{T(c)_i\}_{i=0}^n$, which is the subject of this section, is useful later in obtaining the Schur-Cohn minors from the latter.

CLAIM 3.2.

$$\breve{F}^\sharp(z)_n = \lambda F^\sharp(c)_n \Big|_{c \to z}$$

Proof. Note that

$$\breve{F}^{\sharp}(z)_n = z^n \bar{\breve{F}}\left(\frac{1}{z}\right)_n = \lambda z^n \bar{F}(c)_n \Big|_{c\to z} = \lambda z^n \bar{F}\left(\frac{1-z}{\tau z}\right)_n;$$
$$F^{\sharp}(c)_n \Big|_{c\to z} = (1+\tau c)^n \bar{F}\left(-\frac{c}{1+\tau c}\right)_n \Big|_{c\to z} = z^n \bar{F}\left(\frac{1-z}{\tau z}\right)_n.$$

The claim is thus proven. ∎

THEOREM 3.3. The rows of the complex-q-BT of $\breve{F}(z)$ and the complex-δ-BT of $F(c)$ are related by

$$\breve{T}(z)_i = \begin{cases} \lambda T(c)_i \Big|_{c\to z}, & i = n, n-2, \ldots; \\ \frac{\lambda}{\tau} T(c)_i \Big|_{c\to z}, & i = n-1, n-3, \ldots. \end{cases}$$

Proof. First, using Claim 3.2, note that

$$\breve{T}(z)_n = \lambda T(c)_n \Big|_{c\to z}.$$

Thus, Theorem 3.3 is established for $i = n$. $i = n-1$ may also be established directly. For $i = n-2, n-3, \ldots, 0$, use (2.3) and (3.3). ∎

COROLLARY 3.4.

$$\breve{t}_0^{(i)} = \begin{cases} \frac{\lambda}{\tau^i} t_i^{(i)}, & \text{for } i = n, n-2, \ldots; \\ \frac{\lambda}{\tau^{i+1}} t_i^{(i)}, & \text{for } i = n-1, n-3, \ldots, \end{cases}$$
$$\breve{t}_i^{(i)} = \begin{cases} \frac{\lambda}{\tau^i} \bar{t}_i^{(i)}, & \text{for } i = n, n-2, \ldots; \\ \frac{\lambda}{\tau^{i+1}} \bar{t}_i^{(i)}, & \text{for } i = n-1, n-3, \ldots. \end{cases}$$

Proof. This follows directly from Theorem 3.3. ∎

4. Schur-Cohn Minors for δ-Systems

We now develop quantities that may be considered the analogs of Schur-Cohn minors for δ-system polynomials.

LEMMA 4.1. The relationship between the complex-δ-BT of $F(c)_n \in \Im[c]_n$ and the Schur-Cohn minors $\breve{\Delta}_i$, $i = 1, 2, \ldots, n$, of $\breve{F}(z)_n \in \Im[z]_n$ is

$$\breve{\Delta}_i = \frac{\lambda^2}{2\tau^{2(n-i+1)}} \left[(t_{n-i+1}^{(n-i+1)} \bar{t}_{n-i}^{(n-i)} + \bar{t}_{n-i+1}^{(n-i+1)} t_{n-i}^{(n-i)}) \breve{\Delta}_{i-1} \right.$$
$$\left. - \frac{\lambda^2}{2\tau^{2(n-i+1)}} |t_{n-i+1}^{(n-i+1)} t_{n-i}^{(n-i)}|^2 \breve{\Delta}_{i-2} \right], \ \breve{\Delta}_0 = 1, \ \breve{\Delta}_i = 0, \ i < 0.$$

Proof. Note that, the relationship between the complex-q-BT of $\breve{F}(z)_n$ and its Schur-Cohn minors are given by [25]

$$\breve{\Delta}_i = \frac{1}{2} \left[(\breve{t}_0^{(n-i+1)} \breve{t}_{n-i}^{(n-i)} + \breve{t}_{n-i+1}^{(n-i+1)} \breve{t}_0^{(n-i)}) \breve{\Delta}_{i-1} - \frac{1}{2} |\breve{t}_{n-i+1}^{(n-i+1)} \breve{t}_{n-i}^{(n-i)}|^2 \breve{\Delta}_{i-2} \right],$$

with $\breve{\Delta}_0 = 1$ and $\breve{\Delta}_i = 0$, $i < 0$. Now, the claim follows from Corollary 3.4. ∎

Let

$$D = \operatorname{diag} \left\{ \frac{1}{\tau^n}, \frac{1}{\tau^{n-1}}, \ldots, \frac{1}{\tau} \right\} \in \Re^{n \times n}. \tag{4.1}$$

Then, from Lemma 4.1, $\breve{\Delta}$ in Theorem 2.5 is given by

$$\breve{\Delta} = \lambda^2 \cdot D \cdot \Delta \cdot D \tag{4.2}$$

where

$$\Delta = \begin{bmatrix} \operatorname{Re}[t_n^{(n)} \bar{t}_{n-1}^{(n-1)}] & \frac{\tau}{2}[\bar{t}_{n-1}^{(n-1)} t_{n-2}^{(n-2)}] & 0 & \cdots & 0 \\ \frac{\tau}{2}[t_{n-1}^{(n-1)} \bar{t}_{n-2}^{(n-2)}] & \operatorname{Re}[t_{n-1}^{(n-1)} \bar{t}_{n-2}^{(n-2)}] & \frac{\tau}{2}[\bar{t}_{n-2}^{(n-2)} t_{n-3}^{(n-3)}] & \ddots & 0 \\ 0 & \frac{\tau}{2}[t_{n-2}^{(n-2)} \bar{t}_{n-3}^{(n-3)}] & \operatorname{Re}[t_{n-2}^{(n-2)} \bar{t}_{n-3}^{(n-3)}] & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \operatorname{Re}[t_1^{(1)} \bar{t}_0^{(0)}] \end{bmatrix}. \tag{4.3}$$

Clearly, positive definiteness of $\breve{\Delta}$ and Δ are equivalent statements. Hence, we may consider the principal minors of Δ to be the Schur-Cohn minors of $F(c)$.

DEFINITION 4.1. The Schur-Cohn minors of $F(c) \in \Im[c]_n$ are the principal minors of the tridiagonal Hermitian matrix Δ in (4.3).

Therefore, from Theorem 2.4, we have

THEOREM 4.2. The polynomial $F(c) \in \Im[c]_n$ is stable iff $\Delta_i > 0$, $i = 1, 2, \ldots, n$, where Δ_i is the $(i \times i)$-principal minor of Δ in (4.3).

Remarks.

1. Tridiagonal Hermitian matrices constitute an important class of matrices that have been extensively investigated in matrix theory literature [26]. See also [10].
2. Since the Schur-Cohn minor $\breve{\Delta}_i$ obtained from the complex-q-BT are necessarily proper [10], [25], the Schur-Cohn minors defined above for δ-systems are proper as well.

In terms of the 'scaled' sequence of polynomials $\{U(\zeta)_i\}_{i=0}^n$, Theorem 4.2 may be stated as

COROLLARY 4.3. The polynomial $F(c) \in \Im[c]_n$ is stable iff $\tilde{\Delta}_i > 0$, $i = 1, 2, \ldots, n$, where $\tilde{\Delta}_i$ is the $(i \times i)$-principal minor of

$$\tilde{\Delta} = \begin{bmatrix} -\mathrm{Re}[u_n^{(n)} \bar{u}_{n-1}^{(n-1)}] & -\frac{1}{2}[\bar{u}_{n-1}^{(n-1)} u_{n-2}^{(n-2)}] & 0 & \cdots & 0 \\ -\frac{1}{2}[u_{n-1}^{(n-1)} \bar{u}_{n-2}^{(n-2)}] & -\mathrm{Re}[u_{n-1}^{(n-1)} \bar{u}_{n-2}^{(n-2)}] & -\frac{1}{2}[\bar{u}_{n-2}^{(n-2)} u_{n-3}^{(n-3)}] & \ddots & 0 \\ 0 & -\frac{1}{2}[u_{n-2}^{(n-2)} \bar{u}_{n-3}^{(n-3)}] & -\mathrm{Re}[u_{n-2}^{(n-2)} \bar{u}_{n-3}^{(n-3)}] & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\mathrm{Re}[u_1^{(1)} \bar{u}_0^{(0)}] \end{bmatrix}. \tag{4.3}$$

Proof. Using (3.11), and factoring out the appropriate diagonal matrices, the result immediately follows. ∎

Remark. Again, notice how the use of the 'scaled' sequence simplifies the entries.

5. Algorithm for Checking Stability of 2-D δ-Systems

To check condition II of Theorem 2.2, we may adopt the following approach:
(a) Express $F(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$ as a polynomial in $\Im[c_2]_{n_2}$ so that its coefficients, as well as the corresponding Schur-Cohn minors, are parameterized by $c_1 \in \mathcal{T}_\delta$. Here, we have assumed that $n_1 \geq n_2$; otherwise, the roles of n_1 and n_2 may be interchanged.
(b) Check positivity of each of the Schur-Cohn minors, or positive definiteness of the tridiagonal Hermitian matrix $\Delta \in \Im^{n_2 \times n_2}$, for all $c_1 \in \mathcal{T}_\delta$ (see condition II of Theorem 2.2 and Theorem 4.2). These checks may be simplified by applying a direct extension of Siljak's result [16].

However, construction of the complex-δ-BT and the entries of Δ require complex conjugation of certain entries that are functions of $c_1 \in \mathcal{T}_\delta$. This of course complicates the scheme since $\bar{c}_1 = -c_1/(1 + \tau c_1)$, $\forall c_1 \in \mathcal{T}_\delta$. On the other hand, in dealing with 2-D q-system stability, we have $\bar{z}_1 = 1/z_1$, $\forall z_1 \in \mathcal{T}_q$. This simple relationship has led to stability checking schemes that use the complex forms of tabular forms [10] that incorporate the *polynomial array* method [27]. To circumvent the above difficulty, the algorithm given below uses the real-δ-BT in order to check Theorem 2.3. In the appendix, an easily implementable algorithm that yields

$$G(x, c_2) = G(x)_{n_1}(c_2)_{2n_2} = F(c_1, c_2)F(\bar{c}_1, c_2)\big|_{\substack{c_1 \in \mathcal{T}_\delta \\ x=(c_1+\bar{c}_1)/2}} \in \Re[x]_{n_1}[c_2]_{2n_2} \tag{5.1}$$

is provided. Note that

$$c_1 \in \mathcal{T}_\delta \iff x \in [-2/\tau, 0]. \tag{5.2}$$

Before proceeding, however, it is important to note that tabular methods are useful in checking for no roots to be *outside* the stability region. However, since in typical 2-D stability studies the 2-D transforms are taken with positive powers [9-10], prior to applying the stability check, the following 'preparation' must be done:
(a) Condition I in Theorem 2.3 may be checked by explicitly finding the roots or applying the real-δ-BT to ensure

$$F^\sharp(c_1)(-1/\tau) \doteq (1 + \tau c_1)^{n_1} \bar{F}\left(\frac{-c_1}{1 + \tau c_1}\right)\left(-\frac{1}{\tau}\right) \neq 0, \ \forall c_1 \in \Im \setminus \mathcal{U}_\delta \tag{5.3}$$

(that is, polynomial is reciprocated with respect to c_1).

(b) First form

$$G(x)_{n_1}(c_2)_{2n_2} = \sum_{\ell=0}^{2n_2} g_\ell^{(2n_2)}(x) c_2^\ell \in \Re[x]_{n_1}[c_2]_{2n_2} \quad \text{where} \quad g_\ell^{(2n_2)}(x) = \sum_{k=0}^{n_1} g_{k\ell}^{(2n_2)} x^k \in \Re[x]_{n_1}. \tag{5.4}$$

Here $x \in [-2/\tau, 0]$. Now, condition II in Theorem 2.3 may be checked by applying the real-δ-BT to ensure

$$\begin{aligned} \tilde{G}(x)_{n_1}(c_2)_{2n_2} &\doteq \sum_{\ell=0}^{2n_2} \tilde{g}_\ell^{(2n_2)}(x) c_2^\ell \quad \text{where} \quad \tilde{g}_\ell^{(2n_2)}(x) = \sum_{k=0}^{n_1} \tilde{g}_{k\ell}^{(2n_2)} x^k \in \Re[x]_{n_1} \\ &\doteq G(x)^\sharp(c_2) \\ &= (1+\tau c_2)^{2n_2} G(x)\left(\frac{-c_2}{1+\tau c_2}\right) \neq 0, \ \forall x \in [-2/\tau, 0], \ \forall c_2 \in \Im \setminus \mathcal{U}_\delta \end{aligned} \tag{5.5}$$

(that is, polynomial is reciprocated with respect to c_2). Again, $x \in [-2/\tau, 0]$.

We will hence implicitly assume that the given 2-D δ-polynomial has already been appropriately 'prepared' as above. In addition, the construction of the real-δ-BT for $\tilde{G}(x)(c_2)$ requires ensuring [11]

$$\tilde{g}_0^{(2n_2)}(x) \neq 0 \quad \text{and} \quad \tilde{g}_{2n_2}^{(2n_2)} > 0, \quad \forall x \in [-2/\tau, 0]. \tag{5.6}$$

Violation of the first condition in (5.6) is equivalent to

$$F(c_1)(0) = 0 \quad \text{for some} \quad c_1 \in \mathcal{T}_\delta. \tag{5.7}$$

Assuming, with no loss of generality, $\tilde{g}_{2n_2}^{(2n_2)} > 0$ for some $x \in [-2/\tau, 0]$, violation of the second condition in (5.6) is equivalent to

$$F(c_1)(-1/\tau) = 0 \quad \text{for some} \quad c_1 \in \mathcal{T}_\delta. \tag{5.8}$$

Therefore, each of these violations imply instability. Verifying condition (5.7) must be included in the algorithm. Condition (5.8) is automatically verified when condition I in Theorem 2.3 is checked (see (5.3)).

Then, we have the following

THEOREM 5.1. The 2-D δ-system in (2.7) is stable iff

I. $F(c_1)(-1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and

II. $F(c_1)(0) \neq 0$, $\forall c_1 \in \mathcal{T}_\delta$, and

III. $\Delta_i(0) > 0$, $\forall i = 1, 2, \ldots, 2n_2$, and

IV. $\Delta_{2n_2}(x) > 0$, $\forall x \in [-2/\tau, 0]$, which is satisfied whenever $\Delta_{2n_2}(x) \neq 0$, $\forall x \in [-2/\tau, 0]$, together with condition III.

Here, Δ is the Hermitian matrix mentioned in Theorem 4.2 corresponding to $\tilde{G}(x)(c_2)$ where $x \in [-2/\tau, 0]$.

Conditions I and II in Theorem 5.1 are easy to carry out (they may in fact be verified by explicitly finding the roots). Condition III and IV require construction of the real-δ-BT and the Schur-Cohn minors for which we now develop polynomial arrays [27]. We also provide a scaling scheme so that the numerical reliability of the resulting algorithm is enhanced.

5.1. Polynomial array for entries of real-δ-BT

Express $G(x)(c_2)$ as

$$G(x)(c_2) = \mathbf{x}^{(n_1)^T} \cdot \mathbf{G} \cdot \mathbf{c}_2^{(2n_2)} \tag{5.9}$$

where $\mathbf{x}^{(n_1)} = [x^{n_1}, x^{n_1-1}, \ldots, 1]^T$, $\mathbf{c}_2^{(2n_2)} = [c_2^{2n_2}, c_2^{2n_2-1}, \ldots, 1]^T$, and $\mathbf{G} = \{g_{i,j}\} \in \Re^{(n_1+1)\times(2n_2+1)}$ is the coefficient matrix. Then, it is easy to show that [6]

$$\tilde{G}(x)(c_2) = \mathbf{x}^{(n_1)^T} \cdot \tilde{\mathbf{G}} \cdot \mathbf{c}_2^{(2n_2)} \quad \text{where} \quad \tilde{\mathbf{G}} = \mathbf{G}\tau^{(2n_2)^{-1}}\mathbf{P}^{(2n_2)}\tau^{(2n_2)}. \tag{5.10}$$

Here

$$\begin{aligned} \tau^{(2n_2)} &= \text{diag}\{\tau^{2n_2}, \tau^{2n_2-1}, \ldots, 1\} \in \Re^{(2n_2+1)\times(2n_2+1)}; \\ \mathbf{P}^{(2n_2)} &= \{p_{ij}\} \in \Re^{(2n_2+1)\times(2n_2+1)} \quad \text{where} \quad p_{ij} = (-1)^{2n_2+1-i}\rho_{ij}. \end{aligned} \tag{5.11}$$

The elements ρ_{ij}, which in fact are those of the Pascal's triangle, are given by

$$\rho_{ij} = \begin{cases} 0, & \text{for } i < j; \\ 1, & \text{for } i = j; \\ \rho_{i-1,j-1} + \rho_{i-1,j}, & \text{elsewhere.} \end{cases} \tag{5.12}$$

The real-δ-BT is constructed using the 'scaled' polynomial sequence in (3.11-14). Let

$$\begin{aligned}\tilde{H}(y)(\zeta) &= \tilde{G}(x)(c_2)\,\big|_{\substack{c_2=-\zeta/\tau\\ x=-y/\tau}};\\ H(y)(\zeta) &= \tilde{G}(x)^{\sharp}(c_2)\,\big|_{\substack{c_2=-\zeta/\tau\\ x=-y/\tau}} = G(x)(c_2)\,\big|_{\substack{c_2=-\zeta/\tau\\ x=-y/\tau}}.\end{aligned} \tag{5.13}$$

Note that, $x \in [-2/\tau, 0]$ iff $y \in [0,2]$. Now, using (5.9-12), row $\#2n_2$ and $2n_2-1$ of the corresponding 'scaled' real-δ-BT are given by

$$\begin{aligned}U(y)(\zeta)_{2n_2} &= \sum_{\ell=0}^{2n_2} u_\ell^{(2n_2)}\zeta^\ell = \tilde{H}(y)(\zeta) + H(y)(\zeta)\\ &= \mathbf{y}^{(n_1)^T}\cdot\hat{\boldsymbol{\tau}}^{(n_1)^{-1}}\mathbf{G}\boldsymbol{\tau}^{(2n_2)^{-1}}(\hat{I}^{(2n_2)} + \hat{\mathbf{P}}^{(2n_2)})\cdot\boldsymbol{\zeta}^{(2n_2)};\\ U(y)(\zeta)_{2n_2-1} &= \sum_{\ell=0}^{2n_2-1} u_\ell^{(2n_2-1)}\zeta^\ell = \frac{\tilde{H}(y)(\zeta) - H(y)(\zeta)}{-\zeta/\tau}\\ &= \mathbf{y}^{(n_1)^T}\cdot\hat{\boldsymbol{\tau}}^{(n_1)^{-1}}\tau\mathbf{G}\boldsymbol{\tau}^{(2n_2)^{-1}}(\hat{I}^{(2n_2)} - \hat{\mathbf{P}}^{(2n_2)})\cdot\begin{bmatrix}\boldsymbol{\zeta}^{(2n_2-1)}\\ 0\end{bmatrix},\end{aligned} \tag{5.14}$$

where $\boldsymbol{\zeta}^{(2n_2)} = [\zeta^{2n_2}, \zeta^{2n_2-1}, \ldots, 1]^T$, and

$$\begin{aligned}\hat{\boldsymbol{\tau}}^{(n_1)} &= \mathrm{diag}\{(-\tau)^{n_1}, (-\tau)^{n_1-1}, \ldots, 1\} \in \Re^{(n_1+1)\times(n_1+1)};\\ \hat{I}^{(2n_2)} &= \mathrm{diag}\{(-1)^{2n_2}, (-1)^{2n_2-1}, \ldots, 1\} \in \Re^{(2n_2+1)\times(2n_2+1)};\\ \hat{\mathbf{P}}^{(2n_2)} &= \{\hat{p}_{ij}\} \in \Re^{(2n_2+1)\times(2n_2+1)} \quad \text{where} \quad \hat{p}_{ij} = (-1)^{i+j}\rho_{ij}.\end{aligned} \tag{5.15}$$

Each element of the remaining rows is of the form

$$u_\ell^{(i)}(y) = \frac{n_\ell^{(i)}(y)}{d^{(i)}(y)},\ \ell = 0,1,\ldots,i,\ i = 2n_2, 2n_2-1, \ldots, 0, \tag{5.16}$$

where $n_\ell^{(i)}(y) \in \Re[y]_{\sigma^{(i)}}$ and $d^{(i)}(y) \in \Re[y]_{\varsigma^{(i)}}$. Substituting in (3.12), it is easy to show that, for $\ell = 0,1,\ldots,i$,

$$\begin{aligned}n_\ell^{(i)} &= n_{i+2}^{(i+2)}(n_{\ell-1}^{(i+1)} - 2n_\ell^{(i+1)}) - n_{i+1}^{(i+1)}n_\ell^{(i+2)} + n_{\ell-1}^{(i)}, \quad \text{for} \quad i = 2n_2-2, \ldots, 0;\\ d^{(i)} &= \begin{cases}1, & \text{for } i = 2n_2, 2n_2-1,\\ d^{(i+2)}n_{i+1}^{(i+1)}, & \text{for } i = 2n_2-2, \ldots, 0.\end{cases}\end{aligned} \tag{5.17}$$

Note that $u_\ell^{(2n_2)} = n_\ell^{(2n_2)}$ and $u_\ell^{(2n_2-1)} = n_\ell^{(2n_2-1)}$. Moreover

$$\begin{aligned}\sigma^{(i)} &= \begin{cases}n_1, & \text{for } i = 2n_2, 2n_2-1,\\ \sigma^{(i+2)} + \sigma^{(i+1)}, & \text{for } i = 2n_2-2, \ldots, 0;\end{cases}\\ \varsigma^{(i)} &= \begin{cases}0, & \text{for } i = 2n_2, 2n_2-1,\\ \sigma^{(i)} - n_1, & \text{for } i = 2n_2-2, \ldots, 0.\end{cases}\end{aligned} \tag{5.18}$$

Scaling scheme. Let us scale rows $\#2n_2$ and $\#(2n_2-1)$ so that each coefficient takes values in $[-1,1]$. Correspondingly, for $\ell = 0,1,\ldots,i;\ i = 2n_2, 2n_2-1$, let

$$\begin{aligned} n_\ell^{(i)} &= \lambda^{(i)} \check{n}_\ell^{(i)}; \\ d^{(i)} &= \gamma^{(i)} \check{d}^{(i)}, \end{aligned} \tag{5.19}$$

where $\lambda^{(i)}, \gamma^{(i)} > 0,\ i = 2n_2, 2n_2-1$, are the scaling constants and $[\check{\cdot}]$ denote scaled quantities. Note that

$$\begin{aligned} \frac{n_\ell^{(2n_2)}}{d^{(2n_2)}} &= \frac{\lambda^{(2n_2)}}{\gamma^{(2n_2)}} \frac{\check{n}_\ell^{(2n_2)}}{\check{d}^{(2n_2)}}; \\ \frac{n_\ell^{(2n_2-1)}}{d^{(2n_2-1)}} &= \frac{\lambda^{(2n_2-1)}}{\gamma^{(2n_2-1)}} \frac{\check{n}_\ell^{(2n_2-1)}}{\check{d}^{(2n_2-1)}}. \end{aligned} \tag{5.20}$$

Now, substituting in (5.17-18), we get

$$\begin{aligned} \frac{n_\ell^{(2n_2-2)}}{\lambda^{(2n_2)}\lambda^{(2n_2-1)}} &= \check{n}_{2n_2}^{(2n_2)}(\check{n}_{\ell-1}^{(2n_2-1)} - 2\check{n}_\ell^{(2n_2-1)}) - \check{n}_{2n_2-1}^{(2n_2-1)}\check{n}_\ell^{(2n_2)} + \frac{n_{\ell-1}^{(2n_2-2)}}{\lambda^{(2n_2)}\lambda^{(2n_2-1)}}; \\ \frac{d^{(2n_2-2)}}{\gamma^{(2n_2)}\lambda^{(2n_2-1)}} &= \check{d}^{(2n_2)}\check{n}_{2n_2-1}^{(2n_2-1)}. \end{aligned} \tag{5.21}$$

It can now be seen that, it is only necessary to compute the quantities on the left hand side of (5.21). Then, one may scale these to get

$$\begin{aligned} n_\ell^{(2n_2-2)} &= \lambda^{(2n_2)}\lambda^{(2n_2-1)}\lambda^{(2n_2-2)}\check{n}_\ell^{(2n_2-2)}; \\ d^{(2n_2-2)} &= \gamma^{(2n_2)}\gamma^{(2n_2-2)}\lambda^{(2n_2-1)}\check{d}^{(2n_2-2)}. \end{aligned} \tag{5.22}$$

Note that

$$\frac{n_\ell^{(2n_2-2)}}{d^{(2n_2-2)}} = \frac{\lambda^{(2n_2)}\lambda^{(2n_2-2)}}{\gamma^{(2n_2)}\gamma^{(2n_2-2)}} \frac{\check{n}_\ell^{(2n_2-2)}}{\check{d}^{(2n_2-2)}}. \tag{5.23}$$

Continuing in this manner, the computation of the entries of real-δ-BT may be summarized as follows:

(a) From (5.14), compute $n_\ell^{(i)}, d^{(i)},\ i = 2n_2, 2n_2-1$.

(b) From (5.19), use scaling constants $\lambda^{(i)}, \gamma^{(i)},\ i = 2n_2, 2n_2-1$, to get $\check{n}_\ell^{(i)}, \check{d}^{(i)},\ i = 2n_2, 2n_2-1$.

(c) From (5.21), for $\ell = 0,1,\ldots,i;\ i = 2n_2-2,\ldots,0$, compute

$$\begin{aligned} \frac{n_\ell^{(i)}}{K_n^{(i)}} &= \check{n}_{i+2}^{(i+2)}(\check{n}_{\ell-1}^{(i+1)} - 2\check{n}_\ell^{(i+1)}) - \check{n}_{i+1}^{(i+1)}\check{n}_\ell^{(i+2)} + \frac{n_{\ell-1}^{(i)}}{K_n^{(i)}}; \\ \frac{d^{(i)}}{K_d^{(i)}} &= \check{d}^{(i+2)}\check{n}_{i+1}^{(i+1)}, \end{aligned} \tag{5.24}$$

and use scaling constants $\lambda^{(i)}, \gamma^{(i)},\ i = 2n_2 - 2, \ldots, 0$, to get $\check{n}_\ell^{(i)}, \check{d}^{(i)},\ i = 2n_2 - 2, \ldots, 0$. Here, $K_n^{(i)}$ and $K_d^{(i)}$ are constants.

(d) Notice the relationships

$$\frac{n_\ell^{(i)}}{d^{(i)}} = \begin{cases} \frac{\lambda^{(2n_2)}\lambda^{(2n_2-2)}\ldots\lambda^{(i)}}{\gamma^{(2n_2)}\gamma^{(2n_2-2)}\ldots\gamma^{(i)}} \frac{\check{n}_\ell^{(i)}}{\check{d}^{(i)}}, & \text{for } i = 2n_2, 2n_2-2, \ldots, 0; \\ \frac{\lambda^{(2n_2-1)}\lambda^{(2n_2-3)}\ldots\lambda^{(i)}}{\gamma^{(2n_2-1)}\gamma^{(2n_2-3)}\ldots\gamma^{(i)}} \frac{\check{n}_\ell^{(i)}}{\check{d}^{(i)}}, & \text{for } i = 2n_2-1, 2n_2-3, \ldots, 1. \end{cases} \tag{5.25}$$

5.2. Polynomial array for Schur-Cohn minors

Each Schur-Cohn minor obtained from the table, in general, will be of the form

$$\Delta_i(y) = \frac{N^{(i)}(y)}{D^{(i)}(y)} \in \Re(y),\ i = 1, 2, \ldots, 2n_2, \tag{5.26}$$

where $N^{(i)}(y) \in \Re[y]_{\rho^{(i)}}$ and $D^{(i)}(y) \in \Re[y]_{\varrho^{(i)}}$. From Corollary 4.3, we get

$$\Delta_i(y) = -u_{2n_2-i+1}^{(2n_2-i+1)} u_{2n_2-i}^{(2n_2-i)} \Delta_{i-1} - \frac{1}{4} u_{2n_2-i+1}^{(2n_2-i+1)^2} u_{2n_2-i}^{(2n_2-i)^2} \Delta_{i-2},\ i = 1, 2, \ldots, 2n_2, \tag{5.27}$$

where $\Delta_0 \doteq 1$ and $\Delta_i = 0,\ \forall i < 0$.

Remark. Actually, as in [10], one may show that, for stability determination purposes, only the numerator polynomials of Δ_i need be computed. However, to contain the orders of the resulting polynomials, and hence improve numerically conditioning, we do not recommend this scheme.

Scaling scheme. Due to the scaling of entries of the real-δ-BT, computation of $\Delta_i,\ i = 1, 2, \ldots, 2n_2$, may be modified as follows: Let

$$\begin{aligned} \Delta_1 &= -u_{2n_2}^{(2n_2)} u_{2n_2-1}^{(2n_2-1)} \\ &= -\frac{\lambda^{(2n_2)}\lambda^{(2n_2-1)}}{\gamma^{(n2_2)}\gamma^{(2n_2-1)}} \frac{\check{n}_{2n_2}^{(2n_2)} \check{n}_{2n_2-1}^{(2n_2-1)}}{\check{d}^{(2n_2)}\check{d}^{(2n_2-1)}}. \end{aligned} \tag{5.28}$$

Hence, it is only necessary to compute the quantity

$$\check{\Delta}_1 \doteq -\frac{\check{n}_{2n_2}^{(2n_2)} \check{n}_{2n_2-1}^{(2n_2-1)}}{\check{d}^{(2n_2)}\check{d}^{(2n_2-1)}}. \tag{5.29}$$

Continuing in this manner, the computation of the Schur-Cohn minors may be summarized as follows: From (5.27), for $i = 1, 2, \ldots, 2n_2$, compute

$$\check{\Delta}_i = -\frac{\check{n}_{2n_2-i+1}^{(2n_2-i+1)}\check{n}_{2n_2-i}^{(2n_2-i)}}{\check{d}^{(2n_2-i+1)}\check{d}^{(2n_2-i)}}\left[\check{\Delta}_{i-1} + \frac{1}{4}\frac{\lambda^{(2n_2-i)}}{\gamma^{(2n_2-i)}}\frac{\check{n}_{2n_2-i+1}^{(2n_2-i+1)}\check{n}_{2n_2-i}^{(2n_2-i)}}{\check{d}^{(2n_2-i+1)}\check{d}^{(2n_2-i)}}\check{\Delta}_{i-2}\right], \tag{5.30}$$

where $\check{\Delta}_0 \doteq 1$ and $\check{\Delta}_i = 0$, $\forall i < 0$.

Remark. Note that, since $\Delta_i(y)$ is necessarily a proper polynomial (that is, denominator divides numerator properly with no remainder), and not a rational polynomial (see Remark 2 after Theorem 4.2), it is easy to see that $\check{d}^{(2n_2-i+1)}\check{d}^{(2n_2-i)}$ must divide $\check{n}_{2n_2-i+1}^{(2n_2-i+1)}\check{n}_{2n_2-i}^{(2n_2-i)}$ exactly.

5.3. Algorithm

The following result, which is the basis of the stability checking algorithm, is now obvious from [10] and Theorem 5.1:

THEOREM 5.4. The 2-D δ-system in (2.7) is stable iff

I. $F(c_1, -1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and

II. $F(c_1)(0) \neq 0$, $\forall c_1 \in \mathcal{T}_\delta$, and

III. $\Delta_i(0) > 0$, $\forall i = 1, 2, \ldots, 2n_2$, and

IV. $\Delta_{2n_2}(y) \neq 0$, $\forall y \in [0, 2]$.

The 2-D stability checking algorithm may now be summarized as follows:

GIVEN.

A 2-D δ-polynomial $F(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$. Without any loss of generality, assume that $n_1 \geq n_2$, and express $F(c_1, c_2)$ as $F(c_1)_{n_1}(c_2)_{n_2}$.

STEP I. Condition I of Theorem 5.4:

Apply an explicit root location procedure. If result is satisfactory, proceed; otherwise, system is unstable.

STEP II. Condition II of Theorem 5.4:

Apply an explicit root location procedure. If result is satisfactory, proceed; otherwise, system is unstable.

STEP III.

Form $G(y)(c_2)$ using the algorithm in the appendix; then form $U(y)(\zeta)_{2n_2}$ and $U(y)(\zeta)_{2n_2-1}$ from (5.14). These yield $n_\ell^{(2n_2)}$ and $n_\ell^{(2n_2-1)}$. Of course, $d^{(2n_2)} = d^{(2n_2-1)} = 1$.

From (5.19), obtain $\check{n}_\ell^{(2n_2)}$, $\check{n}_\ell^{(2n_2-1)}$, and the associated scaling constants $\lambda^{(2n_2)}$ and $\lambda^{(2n_2-1)}$. Of course, $\check{d}^{(2n_2)} = \check{d}^{(2n_2-1)} = 1$ and $\gamma^{(2n_2)} = \gamma^{(2n_2-1)} = 1$.

STEP IV. Condition III of Theorem 5.4:

Form $\check{\Delta}_1(y)$ from (5.30) and check whether $\check{\Delta}_1(0) > 0$.

If result is satisfactory, form $\check{n}_\ell^{(2n_2-2)}$ and $\check{d}^{(2n_2-2)}$ and the associated scaling constants $\lambda^{(2n_2-2)}$ and $\gamma^{(2n_2-2)}$ from (5.24). Form $\check{\Delta}_2(y)$ from (5.30) and check whether $\check{\Delta}_2(0) > 0$. If result is satisfactory, proceed likewise until $\check{\Delta}_{2n_2}(0) > 0$ is checked. Note that, this requires checking of only the constant coefficients. If result is satisfactory, proceed; otherwise, if the check fails at any $i = 1, 2, \ldots, 2n_2$, system is unstable.

STEP V. Condition IV of Theorem 5.4:

Apply an explit root location procedure to check whether $\check{\Delta}_{2n_2}(y) \neq 0$, $\forall y \in [0, 2]$.

Remarks. The possible numerical difficulties that may arise in using explit root location procedures may be avoided as follows: (a) Steps I and II may be verified using the real-δ-BT [6], and (b) step V may be verified by the Sturm sequence method.

6. Example

The stability checking algorithm presented in the previous section is now illustrated through an example. Polynomial entries are denoted using a self-explanatory shorthand notation where the highest degree coefficient is written first. Moreover, only four decimal digital on the mantissa are shown.

Consider the 2-D polynomial

$$F(c_1, c_2) = [\, c_1^2 \quad c_1 \quad 1\,] \begin{bmatrix} 1 & 50 & 740 \\ 52 & 2700 & 38480 \\ 740 & 38480 & 547600 \end{bmatrix} \begin{bmatrix} c_2^2 \\ c_2 \\ 1 \end{bmatrix}$$

with the sampling time $\tau = 0.1$ s.

STEP I. Condition I of Theorem 5.4:

By applying an explicit root location procedure, one can show that

$$F(c_1)(-1/\tau) = 340c_1^2 + 16680c_1 + 236800 \neq 0, \ \forall c_1 \in \overline{\mathcal{U}}_\delta.$$

STEP II. Condition II of Theorem 5.4:

By applying an explicit root location procedure, one can show that

$$F(c_1)(0) = 740c_1^2 + 38480c_1 + 547600 \neq 0, \ \forall c_1 \in \mathcal{T}_\delta.$$

STEP III. Using the algorithm in Appendix, we get

$$G(y)(\zeta) = [\, y^2 \quad y \quad 1\,] \begin{bmatrix} 1.2800e+03 & 1.2992e+05 & 5.1904e+06 & 9.6141e+07 & 7.0093e+08 \\ 5.2480e+04 & 5.4011e+06 & 2.1662e+08 & 3.9968e+09 & 2.8738e+10 \\ 5.4760e+05 & 5.6950e+07 & 2.2912e+09 & 4.2143e+10 & 2.9987e+11 \end{bmatrix} \begin{bmatrix} \zeta^2 \\ \zeta \\ 1 \end{bmatrix}.$$

After scaling, rows #4 and #3 are computed as follows:

$$\check{n}_4^{(4)} = [1.2859e-02, -5.0693e-02, 5.1315e-02];$$

$$\check{n}_3^{(4)} = [-7.9833e-02, 3.1991e-01, -3.2798e-01];$$

$$\check{n}_2^{(4)} = [1.9671e-01, -7.9909e-01, 8.2798e-01];$$

$$\check{n}_1^{(4)} = [-2.3375e-01, 9.5836e-01, -1.0000e+00];$$

$$\check{n}_0^{(4)} = [1.1687e-01, -4.7918e-01, 5.0000e-01],$$

with $\lambda^{(4)} = 1.1995e+12$, and

$$\check{n}_3^{(3)} = [-2.4050e-02,\ 9.4053e-02,\ -9.4595e-02];$$
$$\check{n}_2^{(3)} = [1.3044e-01,\ -5.1542e-01,\ 5.2252e-01];$$
$$\check{n}_1^{(3)} = [-2.4703e-01,\ 9.8194e-01,\ -1.0000e+00];$$
$$\check{n}_0^{(3)} = [1.6469e-01,\ -6.5463e-01,\ 6.6667e-01],$$

with $\lambda^{(3)} = 5.3490e+10$. Of course, $\check{d}^{(4)} = \check{d}^{(3)} = 1$ with $\gamma^{(4)} = \gamma^{(3)} = 1$.

STEP IV. Condition III of Theorem 5.4:

We get

$$\check{\Delta}_1 = [3.0926e-04,\ -2.4286e-03,\ 7.2184e-03,\ -9.6217e-03,\ 4.8541e-03].$$

Clearly, $\check{\Delta}_1(0) = 4.8541e-03 > 0$.

Now, row #2 is computed as follows:

$$\check{n}_2^{(2)} = [-8.8619e-03,\ 6.8253e-02,\ -1.9920e-01,\ 2.6099e-01,\ -1.2957e-01];$$
$$\check{n}_1^{(2)} = [3.3584e-02,\ -2.5969e-01,\ 7.6068e-01,\ -1.0000e+00,\ 4.9793e-01];$$
$$\check{n}_0^{(2)} = [-3.3584e-02,\ 2.5969e-01,\ -7.6068e-01,\ 1.0000e+00,\ -4.9793e-01],$$

with $\lambda^{(2)} = 4.2420e-02$. Also,

$$\check{d}^{(2)} = [-2.5424e-01,\ 9.9428e-01,\ -1.0000e+00],$$

with $\gamma^{(2)} = 9.4595e-02$. We get

$$\check{\Delta}_2 = [1.8046e-07,\ -2.8190e-06,\ 1.9343e-05,\ -7.6148e-05,\ 1.8810e-04,$$
$$-2.9857e-04,\ 2.9737e-04,\ -1.6992e-04,\ 4.2654e-05].$$

Clearly, $\check{\Delta}_2(0) = 4.2654e-05 > 0$.

Now, row #1 is computed as follows:

$$\check{n}_1^{(1)} = [2.5168e-03,\ -2.8515e-02,\ 1.3555e-01,\ -3.4597e-01,\ 5.0000e-01,$$
$$-3.8792e-01,\ 1.2623e-01];$$
$$\check{n}_0^{(1)} = [-5.0336e-03,\ 5.7031e-02,\ -2.7110e-01,\ 6.9194e-01,\ -1.0000e+00,$$
$$7.7584e-01,\ -2.5246e-01],$$

with $\lambda^{(1)} = 3.0980e-02$. Also,

$$\check{d}^{(1)} = [-3.3954e-02,\ 2.6151e-01,\ -7.6322e-01,\ 1.0000e+00,\ -4.9646e-01],$$

with $\gamma^{(1)} = 2.6099e-01$. We get

$$\begin{aligned}\check{\Delta}_3 = [&4.0500e-10,\ -9.3525e-09,\ 9.9260e-08,\ -6.4020e-07,\ 2.7947e-06,\\ &-8.6990e-06,\ 1.9797e-05,\ -3.3188e-05,\ 4.0679e-05,\ -3.5552e-05,\\ &2.1029e-05,\ -7.5594e-06,\ 1.2489e-06].\end{aligned}$$

Clearly, $\check{\Delta}_3(0) = 1.2489e-06 > 0$.

Now, row #0 is computed as follows:

$$\begin{aligned}\check{n}_0^{(0)} = [&-1.0487e-04,\ 1.9379e-03,\ -1.6174e-02,\ 8.0291e-02,\ -2.6251e-01,\\ &5.9070e-01,\ -9.2642e-01,\ 1.0000e+00,\ -7.1104e-01,\ 3.0076e-01,\\ &-5.7473e-02],\end{aligned}$$

with $\lambda^{(0)} = 4.4719e-02$. Also,

$$\begin{aligned}\check{d}^{(0)} = [&-6.7946e-04,\ 1.0355e-02,\ -6.9373e-02,\ 2.6679e-01,\ -6.4420e-01,\\ &1.0000e+00,\ -9.7458e-01,\ 5.4519e-01,\ -1.3404e-01],\end{aligned}$$

with $\gamma^{(0)} = 9.4174e-01$. We get

$$\begin{aligned}\check{\Delta}_4 = [&4.3531e-12,\ -1.3058e-10,\ 1.8400e-09,\ -1.6166e-08,\ 9.9118e-08,\\ &-4.4970e-07,\ 1.5618e-06,\ -4.2352e-06,\ 9.0628e-06,\ -1.5355e-05,\\ &2.0530e-05,\ -2.1433e-05,\ 1.7129e-05,\ -1.0130e-05,\ 4.1814e-06,\\ &-1.0762e-06,\ 1.3014e-07].\end{aligned}$$

Clearly, $\check{\Delta}_4(0) = 1.3014e-07 > 0$.

STEP V. Condition IV of Theorem 5.4:

By applying an explicit root location procedure, one can show that

$$\check{\Delta}_4(y) \neq 0,\ \forall y \in [0, 2].$$

Thus, we conclude that $F(c_1, c_2)$ is stable.

7. Conclusion and Final Remarks

In this paper, we have developed an efficient stability checking algorithm applicable for 2-D δ-system characteristic polynomials. Our purpose here is to obtain a direct algorithm due to the possible numerical disadvantages associated with indirect methods that utilize transformation techniques.

In arriving at the algorithm, the following contributions have been made: (a) Tabular method of stability checking applicable for δ-system polynomials possibly possessing complex-valued coefficients, (b) quantities that may be regarded as the Schur-Cohn minors applicable for such systems, and (c) polynomial arrays for computing both table entries and Schur-Cohn minors.

The proposed Schur-Cohn minors lets one use a Siljak-like simplification [16] in the stability check. Although the algorithm utilizes only the real-δ-BT, results regarding the Schur-Cohn minors are in fact valid for the more general complex-valued coefficient case as well.

As in [10], it is possible to develop the algorithm such that only the numerator polynomials of the entries of the real-δ-BT and the Schur-Cohn minors are computed. Then, we do not require polynomial division operations. However, our experience has been that such a scheme is prone to be numerically unreliable. This is mainly due to the explosion of polynomial degree especially in computing the Schur-Cohn minors. To avoid these difficulties and enhance numerical reliability, we have (a) introduced a scaling scheme, and (b) used polynomial division to contain the polynomial degree. The latter is not new; in fact, MJT also uses this. If the user is interested in implementing the algorithm using PRO-MATLAB [28], these polynomial division operations may be conveniently performed using the routine `deconv`.

We believe that a suitable scaling strategy can improve the numerical reliability of the MJT as well. The authors are currently looking into this.

The algorithm developed is easily implementable on a computer. The authors have

implemented it via a C-language routine that the interested reader may request from the second author.

Acknowledgement

The authors are indebted to Professor Eliahu I. Jury for many helpful discussions and the reviewers for their careful examination of the manuscript and constructive suggestions. The first author's research work was partially supported by the Office of Naval Research (ONR) through the grant N00014-94-1-0454. This support is gratefully acknowledged.

References

1. R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1990.
2. R. Vijayan, H.V. Poor, J.B. Moore, and G.C. Goodwin, "A Levinson-type algorithm for modeling fast-sampled data," *IEEE Transactions on Automatic Control,* vol. 36, pp. 314-321, Mar. 1991.
3. C.B. Soh, "Robust stability of discrete-time systems using delta operators," *IEEE Transactions on Automatic Control,* vol. 36, pp. 377-380, 1991.
4. G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE,* vol. 80, pp. 240-259, 1992.
5. G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Transactions on Signal Processing,* vol. 41, pp. 629-637, 1993.
6. K. Premaratne and E.I. Jury, "Tabular method for determining root distribution of delta operator formulated polynomials," *IEEE Transactions on Automatic Control,* vol. 39, pp..352-355, 1994.
7. K. Premaratne, R. Salvi, N.R. Habib, and J.P. Le Gall, "Delta-operator formulated discrete-time equivalents of continuous-time systems," *IEEE Transactions on Automatic Control,* vol. 39, pp. 581-585, 1994.
8. G. Likourezos, "Prolog to 'High-speed digital signal processing and control'," *Proceedings of the IEEE,* vol. 80, 238-239, 1992.
9. E.I. Jury, "Stability of multidimensional systems and other related problems," Chapter 3 in *Multidimensional Systems, Techniques, and Applications,* New York, NY: Marcel Dekkar, 1986.
10. K. Premaratne, "Stability determination of two-dimensional discrete-time systems," *Multidimensional Systems and Signal Processing,* vol. 4, pp. 331-354, 1993.
11. Y. Bistritz, "Zero location with respect to the unit circle of discrete-time linear system polynomials," *Proceedings of the IEEE,* vol. 72, pp. 1131-1142, 1984.
12. M. Mansour, "Stability and robust stability of discrete-time systems in the δ-transform," *Fundamentals of Discrete-Time Systems: A Tribute to Professor Eliahu I. Jury,* pp. 133-140, Albuquerque, NM: TSI Press, 1993.
13. R. DeCarlo, J. Murray, and R. Saeks, "Multivariable Nyquist theory," *International Journal of Control,* vol. 25, pp. 657-675, 1977.
14. G. Gu and E.B. Lee, "A numerical algorithm for stability testing of 2-D recursive digital filters," *IEEE Transactions on Circuits and Systems,* vol. 37, pp. 135-138, 1990.
15. E.I. Jury, "Modified stability table for 2-D digital filters," *IEEE Transactions on Circuits and Systems,* vol. 35, pp. 116-119, 1988; "Addendum to 'Modified stability table for 2-D digital filters'," Department Electrical and Computer Engineering, University of Miami, Coral Gables, FL, 1987.
16. D.D. Siljak, "Stability criteria for two-dimensional polynomials," *IEEE Transactions on Circuits and Systems,* vol. 22, pp. 185-189, 1975.
17. E.I. Jury, "A note on the modified stability table for linear discrete-time systems," *IEEE Transactions on Circuits and Systems,* vol. 38, pp. 221-223, 1991.
18. E.I. Jury, *Theory and Application of the z-Transform Method,* John Wiley & Sons: New York, NY, 1964.
19. Y. Bistritz, "A circular stability test for general polynomials," *Systems and Control Letters,* vol. 7, pp. 89-97, 1986.
20. T.S. Huang, "Stability of two-dimensional recursive filters," *IEEE Transactions on Automatic Control,* vol. 20, pp. 158-183, 1973.
21. N.K. Bose, "Implementation of a new stability test for two-dimensional filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 25, pp. 117-120, 1977.
22. B.M. Karan and M.C. Srivastava, "A new stability test for 2-D filters," *IEEE Transactions on Circuits and Systems,* vol. 33, pp. 807-809, 1986.
23. A. Cohn, "Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise," *Math. Z.,* vol. 14-15, pp. 110-148, 1914.
24. I. Schur, "Über Potenzreihen die in Inners des Einheitskreises besehränkt sind," *J. Für Math.,* vol. 147, pp. 205-232, 1917.
25. K. Premaratne and E.I. Jury, "On the Bistritz tabular form and its relationship with the Schur-Cohn minors and inner determinants," *Journal of the Franklin Institute,* vol. 330, pp. 165-182, 1993.
26. G.H. Golub and C.F. Van Loan, *Matrix Computations,* Baltimore, MD: John Hopkins University Press, 1983.
27. X. Hu and E.I. Jury, "On two-dimensional filter stability test," to appear in *IEEE Transactions on Circuits and Systems,* 1993.
28. *PRO-MATLAB User's Guide,* South Natick, MA: The MathWorks Inc., 1991.

Appendix. Algorithm to obtain $G(x)_{n_1}(c_2)_{2n_2}$ from $F(c_1)_{n_1}(c_2)_{n_2}$

Given

$$F(c_1)_{n_1}(c_2)_{n_2} = \sum_{\ell=0}^{n_2} f_\ell(c_1)\cdot c_2^\ell \quad \text{where} \quad f_\ell(c_1) = \sum_{k=0}^{n_1} f_{k,\ell}\cdot c_1^k,\ c_1 \in \mathcal{T}_\delta, \tag{a.1}$$

we now develop an algorithm that yields

$$G(x)_{n_1}(c_2)_{2n_2} \doteq \sum_{j=0}^{2n_2} g_j(x)\cdot c_2^j = F(c_1)_{n_1}(c_2)_{n_2}\cdot F(\bar{c}_1)_{n_1}(c_2)_{n_2},\ c_1 \in \mathcal{T}_\delta. \tag{a.2}$$

First, we see

$$\begin{aligned} G(x)(c_2) &= \sum_{\ell=0}^{n_2}\sum_{j=0}^{n_2} f_\ell(c_1) f_j(\bar{c}_1)\cdot c_2^{\ell+j} \\ &= \sum_{\ell=0}^{n_2}\sum_{j=\ell}^{n_2+\ell} f_\ell(c_1) f_{j-\ell}(\bar{c}_1)\cdot c_2^j = \sum_{j=0}^{2n_2}\sum_{\ell=0}^{j} f_\ell(c_1) f_{j-\ell}(\bar{c}_1)\cdot c_2^j \end{aligned} \tag{a.3}$$

(quantities with negative subscripts are taken to be zero). Hence, comparing (a.2-3), we get

$$\begin{aligned} g_j(x) &= \sum_{\ell=0}^{j} f_\ell(c_1) f_{j-\ell}(\bar{c}_1) = \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1}\sum_{i=0}^{n_1} f_{k,\ell} f_{i,j-\ell}\cdot c_1^k \bar{c}_1^i\right] \\ &= \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1} f_{k,\ell} f_{k,j-\ell}\cdot (c_1\bar{c}_1)^k + \sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{n_1} f_{k,\ell} f_{i,j-\ell}\cdot c_1^k \bar{c}_1^i\right] \\ &= \sum_{\ell=0}^{j}\sum_{k=0}^{n_1} f_{k,\ell} f_{k,j-\ell}\cdot (c_1\bar{c}_1)^k + X \end{aligned} \tag{a.4}$$

where

$$\begin{aligned} X &= \sum_{\ell=0}^{j}\sum_{k=0}^{n_1}\left[\sum_{\substack{i=0\\ i\neq k}}^{k} f_{k,\ell} f_{i,j-\ell}\cdot c_1^k \bar{c}_1^i + \sum_{\substack{i=k\\ i\neq k}}^{n_1} f_{k,\ell} f_{i,j-\ell}\cdot c_1^k \bar{c}_1^i\right] \\ &= \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{k} f_{k,\ell} f_{i,j-\ell}\cdot c_1^k \bar{c}_1^i + \sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{k} f_{i,\ell} f_{k,j-\ell}\cdot c_1^i \bar{c}_1^k\right] \\ &= \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{k}\left(f_{k,\ell} f_{i,j-\ell}\cdot c_1^{k-i} + f_{i,\ell} f_{k,j-\ell}\cdot \bar{c}_1^{k-i}\right)\cdot (c_1\bar{c}_1)^i\right] \\ &= \sum_{k=0}^{n_1}\sum_{i=0}^{k}(c_1\bar{c}_1)^i \sum_{\ell=0}^{j}\left[f_{k,\ell} f_{i,j-\ell}\cdot c_1^{k-i} + f_{i,j-\ell} f_{k,\ell}\cdot \bar{c}_1^{k-i}\right]. \end{aligned} \tag{a.5}$$

Let us use the notation

$$c_1^{(n)} = \frac{c_1^n + \bar{c}_1^n}{2},\ c_1 \in \mathcal{T}_\delta,\ n = 0, 1, \ldots. \tag{a.6}$$

Noting that, for $c_1 \in \mathcal{T}_\delta$,

$$\bar{c}_1 = -\frac{c_1}{1 + \tau c_1}, \tag{a.7}$$

it is easy to show that

$$c_1 \bar{c}_1 = -\frac{2}{\tau} c_1^{(1)}. \tag{a.8}$$

Substituting in (a.5), we get

$$X = \sum_{\ell=0}^{j} \left[\sum_{k=0}^{n_1} \sum_{i=0}^{k-1} 2 f_{k,\ell} f_{i,j-\ell} \left(\frac{-2}{\tau} \right)^i \cdot c_1^{(1)^i} c_1^{(k-i)} \right]. \tag{a.9}$$

Substituting in (a.4), we get

$$g_j(x) = \sum_{\ell=0}^{j} \sum_{k=0}^{n_1} \left[f_{k,\ell} f_{k,j-\ell} \left(\frac{-2}{\tau} \right)^k \cdot c_1^{(1)^k} + \sum_{i=0}^{k-1} 2 f_{k,\ell} f_{i,j-\ell} \left(\frac{-2}{\tau} \right)^i \cdot c_1^{(1)^i} c_1^{(k-i)} \right]. \tag{a.10}$$

Now, in order to develop the algorithm, we need a recursive procedure to compute $c_1^{(n)}$, $n = 0, 1, \ldots$. To proceed, we note that

$$\begin{aligned} c_1^{(n)} &= \frac{(c_1 + \bar{c}_1)(c_1^{n-1} + \bar{c}_1^{n-1}) - c_1 \bar{c}_1 (c_1^{n-2} + \bar{c}_1^{n-2})}{2} \\ &= 2c_1^{(1)} \left(c_1^{(n-1)} + \frac{1}{\tau} c_1^{(n-2)} \right),\ n = 2, 3, \ldots. \end{aligned} \tag{a.11}$$

Let

$$c_1^{(n)} = \sum_{i=0}^{n} c_{1,i}^{(n)} x^i \tag{a.12}$$

where

$$c_1^{(1)} \doteq x. \tag{a.13}$$

Remark. Note that

$$c_1^{(0)} = 1. \tag{a.14}$$

Substituting (a.12) in (a.11), and equating similar coefficients, we get

$$c_{1,i}^{(n)} = 2 \left(c_{1,i-1}^{(n-1)} + \frac{1}{\tau} c_{1,i-1}^{(n-2)} \right),\ i = 0, \ldots, n,\ n = 2, 3, \ldots. \tag{a.15}$$

For instance, $c_1^{(n)}$, $n = 0, 1, \ldots, 5$, may be conveniently obtained from

$$\begin{bmatrix} c_1^{(0)} \\ c_1^{(1)} \\ c_1^{(2)} \\ c_1^{(3)} \\ c_1^{(4)} \\ c_1^{(5)} \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 2/\tau & 2 & & & \\ 0 & 0 & 6/\tau & 4 & & \\ 0 & 0 & 4/\tau^2 & 16/\tau & 8 & \\ 0 & 0 & 0 & 20/\tau^2 & 40/\tau & 16 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{bmatrix}.$$

Two-Dimensional Delta-Operator Formulated Discrete-Time Systems: State-Space Realization and Its Coefficient Sensitivity Properties

K. Premaratne, *Senior Member, IEEE,* J. Suarez, *Student Member, IEEE,* M.M. Ekanayake, *Student Member, IEEE,* and P.H. Bauer, *Member, IEEE*

Abstract. **Recently, delta-operator based implementation of one-dimensional discrete-time systems has been the focus of considerable research activity. This is due mainly to its superior finite wordlength properties and the possibility of providing a unified treatment of both continuous- and discrete-time systems. In this paper, we investigate the delta-operator formulated implementation of two-dimensional discrete-time systems. For this purpose, a local state-space realization that is analogous to the Roesser model is introduced. Reachability and observability gramians and the notion of a balanced realization for such a model are defined. Coefficient sensitivity properties of the resulting implementations, under both fixed- and floating-point arithmetic, are also carried out.**

Manuscript received—

K. Premaratne, J. Suarez, and M.M. Ekanayake are with the Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124, USA. P.H. Bauer is with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA.

K.P. and P.H.B. gratefully acknowledge the support provided by the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

I. Introduction

Current interest in δ-systems is due mainly to two reasons: (a) δ-systems provide possess superior roundoff noise propagation (Li and Gevers 1993) and coefficient sensitivity (Li and Gevers 1990, Premaratne, et. al. 1994) properties, and (b) the δ-operator makes it possible to treat both continuous- and discrete-time systems in a unified manner since it yields the differential operator as a limiting case (Middleton and Goodwin 1990, Vijayan, et. al. 1991).

Hence, implementation of two-dimensional (2-D) and multi-dimensional (m-D) using the δ-operator can be expected to provide digital filters that perform better in a shorter wordlength environment. If this is the case, such implementations can find widespread use in real-time applications, such as, use of narrow bandwidth filters with high sampling rates where traditional q-operator implementations perform poorly (Likourezos 1991).

With the above in mind, research directed towards developing models for 2-D and m-D δ-systems is warranted. This paper presents a local state-space model that is completely analogous to the well known q-operator Roesser model (Roesser 1975). We also define the notions of gramians and balanced realization, and with these tools in hand, investigate coefficient sensitivity properties of this model. Indeed, implementation of 2-D and m-D systems using this *Roesser δ-model,* under mild conditions, is shown to provide superior coefficient sensitivity properties compared with the more conventional implementation of *Roesser q-model.* As usual, for notational simplicity, we concentrate only on the 2-D case, the extension to the m-D case being quite straight-forward.

The paper is organized as follows: Section II contains nomenclature and some preliminary and a brief review of relevant results. Section III contains the development of the Roesser δ-model and some important system theoretic notions. In particular, after establishing the connection between the gramians of one-dimensional (1-D) q- and δ-systems, we define the notion of gramians for 2-D δ-systems. The relationship between these and those corresponding to 2-D q-systems, gramian computation for separable systems, and the notion of a balanced (BL) realization are then presented. Investigation of the coefficient sensitivity of the δ-model is in Section IV. Addresing the more general multi-input multi-output (MIMO) case, for this purpose, two sensitivity measures applicable for fixed-point (FXP) and floating-point (FLP) arithmetic schemes are proposed. In the case of FXP arithmetic, we show that, BL realizations possess excellent coefficient sensitivity properties. Section V contains an example. Section VI is reserved for concluding remarks.

II. Nomenclature and Preliminaries

2.2. Nomenclature

$\Re$, $\Im$, $\aleph$	Real numbers, complex numbers, and nonnegative integers, respectively.
$\Re^{q\times p}$, $\Im^{q\times p}$	Set of matrices of size $q \times p$ over $\Re$ and $\Im$, respectively.
$\Re[w]_n$, $\Im[w]_n$	Set of univariate polynomials of degree n (with respect to the indeterminate $w \in \Im$) over $\Re$ and $\Im$, respectively.
$\Re(w)_n$	Set of rational univariate polynomials of degree n (with respect to the indeterminate $w \in \Im$) over $\Re$.
$\Re[w_h]_{n_h}[w_v]_{n_v}$	Set of bivariate polynomials of relative degrees n_h and n_v (with respect to the indeterminates $w_h \in \Im$ and $w_v \in \Im$, respectively) over $\Re$.
$\Re(w_h)_{n_h}(w_v)_{n_v}$	Set of rational bivariate polynomials of relative degrees n_h and n_v (with respect to the indeterminates $w_h \in \Im$ and $w_v \in \Im$, respectively) over $\Re$.
I_n	Unit matrix of size $n \times n$.
$\{a_{ij}\}$	Elements of matrix A.
A^*, A^T	Complex conjugate transpose and transpose of matrix A.
trace$[A]$, $\lambda_i[A]$	Trace and i-th eigenvalue of matrix A.
$\oplus$, $\otimes$	Matrix Kronecker sum and product operators, respectively.
$\mathbf{e}_i^{(n)}$	Unit vector in $\Re^n$ with 1 on the i-th row.
$E_{i,j}^{q\times p}$	$\mathbf{e}_i^{(q)}\mathbf{e}_j^{(p)^*} \in \Re^{q\times p}$.
$\overline{U}_{q\times p}$	$\sum_{i=1}^{q}\sum_{j=1}^{p} E_{i,j}^{(q\times p)} \otimes E_{i,j}^{(q\times p)} \in \Re^{q^2\times p^2}$.
$\|A\|_F$	Fröbenius norm of A.

In dealing with q-systems, the conventional indeterminate z (with or without a subscript) is used; for δ-systems, we use the indeterminate c (with or without a subscript). For 1-D systems, the transformation relationship between corresponding q- and δ-systems is

$$\delta = \frac{q-1}{\tau} \Longleftrightarrow c = \frac{z-1}{\tau},$$

where τ is a positive real constant, usually the sampling time.

For 2-D systems, the subscripts h and v denote the horizontally propagating (h.p.) and vertically propagating (v.p.) subsystems of the corresponding Roesser local state-space models.

τ_h, τ_v	Positive real constants denoting the 'sampling times' along the h.p. v.p. directions, respectively.
ξ, Ξ	$\tau_h I_{n_h} \oplus \tau_v I_{n_v} \in \Re^{n\times n}$, $\tau_h I_{n_h q} \oplus \tau_v I_{n_v q} \in \Re^{nq\times nq}$.
n_h, n_v	Integer valued symbols denoting the sizes of the realization of the h.p. and v.p. subsystems, respectively.
n	$n_h + n_v$.
I_c, I_z	$c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n\times n}$, $z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$.

For 2-D systems, the transformation relationships between corresponding q- and δ-systems is

$$\delta_h = \frac{q_h - 1}{\tau_h} \Longleftrightarrow c_h = \frac{z_h - 1}{\tau_h};$$
$$\delta_v = \frac{q_v - 1}{\tau_v} \Longleftrightarrow c_v = \frac{z_v - 1}{\tau_v}.$$

q-system quantity analogous to its corresponding δ-system quantity $[\cdot]$ is denoted by $\hat{[\cdot]}$; for example, state-space realization of a given discrete-time system is either $\{A, B, C, D\}$ if implemented based on the δ-operator or $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ if implemented based on the q-operator.

$H(c_h, c_v)\|_{\mathbf{c}\to\mathbf{z}}$	$H(c_h, c_v)\|_{\substack{c_h=(z_h-1)/\tau_h \\ c_v=(z_v-1)/\tau_v}}$.
$G(z_h, z_v)\|_{\mathbf{z}\to\mathbf{c}}$	$G(z_h, z_v)\|_{\substack{z_h=1+\tau_h c_h \\ z_v=1+\tau_v c_v}}$.

Stability studies of 2-D q- and δ-systems involve the following regions:

$\mathcal{U}_q^2, \overline{\mathcal{U}}_q^2, \mathcal{T}_q^2$	$\{(z_h, z_v) \in \Im^2 : \|z_h\| < 1, \|z_v\| < 1\}$, $\{(z_h, z_v) \in \Im^2 : \|z_h\| \le 1, \|z_v\| \le 1\}$, $\{(z_h, z_v) \in \Im^2 : \|z_h\| = 1, \|z_v\| = 1\}$.
$\mathcal{U}_\delta^2, \overline{\mathcal{U}}_\delta^2, \mathcal{T}_\delta^2$	$\{(c_h, c_v) \in \Im^2 : \|c_h + 1/\tau_h\| < 1/\tau_h, \|c_v + 1/\tau_v\| < 1\}$, $\{(c_h, c_v) \in \Im^2 : \|c_h + 1/\tau_h\| \le 1/\tau_h, \|c_v + 1/\tau_v\| \le 1\}$, $\{(c_h, c_v) \in \Im^2 : \|c_h + 1/\tau_h\| = 1/\tau_h, \|c_v + 1/\tau_v\| = 1\}$.

A q-system polynomial with all its roots in $\mathcal{U}_q$ (for the 1-D case) or $\mathcal{U}_q^2$ (for the 2-D case) is said to be *stable*. The corresponding regions for a δ-system polynomial are $\mathcal{U}_\delta$ (for the 1-D case) and $\mathcal{U}_\delta^2$ (for the 2-D case), respectively.

2.2. Preliminaries

First, we provide a brief introduction to the Roesser local state-space model applicable to 2-D q-operator based discrete-time systems (Roesser 1975).

DEFINITION 2.1. The following partial ordering in $\aleph^2$ is used:

$$\begin{aligned}
(h,k) \leq (i,j) &\iff h \leq i \quad \text{and} \quad k \leq j; \\
(h,k) = (i,j) &\iff h = i \quad \text{and} \quad k = j; \\
(h,k) < (i,j) &\iff (h,k) \leq (i,j) \quad \text{and} \quad (h,k) \neq (i,j).
\end{aligned}$$

The 2-D dynamical systems under consideration are assumed to be linear, shift-invariant, and strictly causal. Moreover, they are taken to be modeled by a set of first-order vector difference equations over $\Re$. Given such a p-input and q-output 2-D system, its n_hh-n_vv Roesser local s.s. model takes the following form (Roesser 1975):

$$\begin{aligned}
\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} \hat{A}^{(1)} & \hat{A}^{(2)} \\ \hat{A}^{(3)} & \hat{A}^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} \hat{B}^{(1)} \\ \hat{B}^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [\hat{A}] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [\hat{B}]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= [\,\hat{C}^{(1)} \quad \hat{C}^{(2)}\,] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [\hat{D}]\mathbf{u}(i,j) \\
&\doteq [\hat{C}] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [\hat{D}]\mathbf{u}(i,j),
\end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. Also, $\hat{A}^{(1)} \in \Re^{n_h \times n_h}$, $\hat{A}^{(2)} \in \Re^{n_h \times n_v}$, $\hat{A}^{(3)} \in \Re^{n_v \times n_h}$, $\hat{A}^{(4)} \in \Re^{n_v \times n_v}$, $\hat{B}^{(1)} \in \Re^{n_h \times p}$, $\hat{B}^{(2)} \in \Re^{n_v \times p}$, $\hat{C}^{(1)} \in \Re^{q \times n_h}$, $\hat{C}^{(2)} \in \Re^{q \times n_v}$, $\hat{D} \in \Re^{q \times p}$, and $(i,j) \in \aleph^2$. The operators $q_h[\cdot]$ and $q_v[\cdot]$ denote

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

The s.s. model in (2.1) is typically denoted by the quadruple $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$. The corresponding 2-D characteristic equation and the 2-D transfer function it realizes are given by

$$\begin{aligned}
\det[I_z - \hat{A}] &= \det[z_h I_{n_h} \oplus z_v I_{n_v} - \hat{A}] \in \Re[z_h]_{n_h}[z_v]_{n_v}; \\
\hat{H}(z_h, z_v) &= \hat{C}(I_z - \hat{A})^{-1}\hat{B} + \hat{D} \in \Re(z_h)_{n_h}(z_v)_{n_v},
\end{aligned} \tag{2.3}$$

where $z_h, z_v \in \Im$. In the literature, $\mathbf{x}^h$ and $\mathbf{x}^v$ are referred to as the *horizontally propagating (h.p.)* and *vertically propagating (v.p.)* local state vectors of the s.s. model $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$.

Assuming no nonessential singularities of the second kind on $\mathcal{T}_q^2$, for BIBO stability of the s.s. model above, it is necessary and sufficient that (see Jury 1986, and references therein)

$$\det[I_z - \hat{A}] \neq 0, \ \forall (z_h, z_v) \in \overline{\mathcal{U}}_q^2. \tag{2.4}$$

For investigating coefficient sensitivity properties, we will use certain relationships encountered in Kronecker products and matrix differentiation. The following are from Brewer (1978).

The derivative of $A = \{a_{ij}\} \in \Re^{q \times p}$ with respect to $b \in \Re$ is

$$\frac{\partial A}{\partial b} = \frac{\partial a_{ij}}{\partial b} \in \Re^{q \times p}. \tag{2.5}$$

Hence

$$\left\| \frac{\partial A}{\partial b} \right\|_F^2 = \sum_{i=1}^{q} \sum_{j=1}^{p} \left(\frac{\partial a_{ij}}{\partial b} \right)^2. \tag{2.6}$$

The derivative of $A = \{a_{ij}\} \in \Re^{q \times p}$ with respect to $B = \{b_{k\ell} \in \Re^{s \times r}$ is the partitioned matrix whose (k, ℓ)-th partition is $\partial A / \partial b_{k\ell}$, that is,

$$\frac{\partial A}{\partial B} = \begin{bmatrix} \frac{\partial A}{\partial b_{11}} & \cdots & \frac{\partial A}{\partial b_{1r}} \\ \vdots & \ddots & \vdots \\ \frac{\partial A}{\partial b_{s1}} & \cdots & \frac{\partial A}{\partial b_{sr}} \end{bmatrix} \in \Re^{qs \times pr}. \tag{2.7}$$

Hence

$$\left\| \frac{\partial A}{\partial B} \right\|_F^2 = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{s} \sum_{\ell=1}^{r} \left(\frac{\partial a_{ij}}{\partial b_{k\ell}} \right)^2. \tag{2.8}$$

III. State-Space Model for δ-Operator Implementation

3.1. Local state-space model

To exploit the superior finite wordlength properties of δ-operator implementations, analogous to the 1-D case, let us define the operators $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as follows:

$$\begin{aligned}\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j)-\mathbf{x}(i,j)}{\tau_h} = \frac{q_h[\mathbf{x}](i,j)-\mathbf{x}(i,j)}{\tau_h};\\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1)-\mathbf{x}(i,j)}{\tau_v} = \frac{q_v[\mathbf{x}](i,j)-\mathbf{x}(i,j)}{\tau_v},\end{aligned} \tag{3.1}$$

where τ_h and τ_v are two positive real numbers. Hence, the following relationships are applicable:

$$\begin{aligned}\delta_h &= \frac{q_h-1}{\tau_h} \iff q_h = 1+\tau_h\delta_h;\\ \delta_v &= \frac{q_v-1}{\tau_v} \iff q_v = 1+\tau_v\delta_v.\end{aligned} \tag{3.2}$$

Remark. When τ_h and τ_v are the 'sampling times' corresponding to the horizontal and vertical spatial directions, the operators δ_h and δ_v in fact provide the first-order forward Euler approximants of the derivatives along their corresponding directions. When $\tau_h \to 0$ and $\tau_v \to 0$, the operators δ_h and δ_v yield these derivatives. In the 1-D case, this is the reason for the possibility of a unified treatment of both continuous- and discrete-time systems (Middleton and Goodwin 1990).

With (3.2) in mind, we get

$$\begin{aligned}\begin{bmatrix}\delta_h[\mathbf{x}^h](i,j)\\ \delta_v[\mathbf{x}^v](i,j)\end{bmatrix} &= \xi^{-1}\begin{bmatrix}(q_h-1)I_{n_h} & \mathbf{0}\\ \mathbf{0} & (q_v-1)I_{n_v}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix}\\ \iff \begin{bmatrix}q_h[\mathbf{x}^h](i,j)\\ q_v[\mathbf{x}^v](i,j)\end{bmatrix} &= I_n + \xi\begin{bmatrix}\delta_h I_{n_h} & \mathbf{0}\\ \mathbf{0} & \delta_v I_{n_v}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix}.\end{aligned} \tag{3.3}$$

Here,

$$\xi = [\tau_h I_{n_h} \oplus \tau_v I_{n_v}] \in \Re^{n\times n}. \tag{3.4}$$

Using (3.3) in (2.1), it is easy to get the following:

$$\begin{aligned}\begin{bmatrix}\delta_h[\mathbf{x}^h](i,j)\\ \delta_v[\mathbf{x}^v](i,j)\end{bmatrix} &= \begin{bmatrix}A^{(1)} & A^{(2)}\\ A^{(3)} & A^{(4)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + \begin{bmatrix}B^{(1)}\\ B^{(2)}\end{bmatrix}\mathbf{u}(i,j)\\ &\doteq [A]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [B]\mathbf{u}(i,j);\\ \mathbf{y}(i,j) &= [\,C^{(1)} \quad C^{(2)}\,]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [D]\mathbf{u}(i,j)\\ &\doteq [C]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [D]\mathbf{u}(i,j).\end{aligned} \tag{3.5}$$

In addition, as opposed to its corresponding q-operator implementation, in a δ-operator implementation, one must perform the following computations:

$$\begin{aligned} \mathbf{x}^h(i+1,j) &= \mathbf{x}^h(i,j) + \tau_h \cdot \delta_h[\mathbf{x}^h](i,j); \\ \mathbf{x}^v(i,j+1) &= \mathbf{x}^v(i,j) + \tau_v \cdot \delta_v[\mathbf{x}^v](i,j). \end{aligned} \tag{3.6}$$

Here,

$$\begin{aligned} A = \xi^{-1}(\hat{A} - I_n) &\iff \hat{A} = I_n + \xi A; \\ B = \xi^{-1}\hat{B} &\iff \hat{B} = \xi B; \\ C = \hat{C} &\iff \hat{C} = C; \\ D = \hat{D} &\iff \hat{D} = D. \end{aligned} \tag{3.7}$$

The size of each submatrix in (3.5) is equal to the corresponding submatrix of the realization in (2.1). In the sequel, the realization $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ in (1.1) will be referred to as the *q-model,* while the realization $\{A, B, C, D\}$ in (3.5) will be referred to as the *δ-model.*

3.2. Properties of the δ-model

The general response equation of the δ-model may be derived in a manner that is exactly analogous to that in Roesser (1975). Hence, in what follows, only the salient results are given, detailed derivations being omitted for the sake of brevity.

The general response of the δ-model is given by

$$\begin{aligned} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} &= \sum_{k=0}^{j} A^{i,j-k} \begin{bmatrix} \mathbf{x}^h(0,k) \\ 0 \end{bmatrix} + \sum_{h=0}^{i} A^{i-j,j} \begin{bmatrix} 0 \\ \mathbf{x}^v(h,0) \end{bmatrix} \\ &\quad + \sum_{(0,0)\le(h,k)<(i,j)} \left(A^{i-h-1,j-k}\xi \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i-h,j-k-1}\xi \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix} \right) \mathbf{u}(h,k); \\ \mathbf{y}(i,j) &= [\, C^{(1)} \quad C^{(2)} \,] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j). \end{aligned} \tag{3.8}$$

Here, $A^{i,j}$ refers to the *transition matrix* of the δ-model. With the partial ordering in $\aleph^2$ agreed upon previously (Definition 2.1), it may be recursively computed as follows:

$$A^{i,j} = \begin{cases} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, & \text{for } (i,j) < (0,0); \\ \begin{bmatrix} I_{n_h} & 0 \\ 0 & I_{n_v} \end{bmatrix}, & \text{for } (i,j) = (0,0); \\ \begin{bmatrix} I_{n_h} & 0 \\ 0 & 0 \end{bmatrix} + \xi \begin{bmatrix} A^{(1)} & A^{(2)} \\ 0 & 0 \end{bmatrix}, & \text{for } (i,j) = (1,0); \\ \begin{bmatrix} 0 & 0 \\ 0 & I_{n_v} \end{bmatrix} + \xi \begin{bmatrix} 0 & 0 \\ A^{(3)} & A^{(4)} \end{bmatrix}, & \text{for } (i,j) = (0,1); \\ A^{1,0}A^{i-1,j} + A^{0,1}A^{i,j-1}, & \text{elsewhere.} \end{cases} \tag{3.9}$$

Remarks.

1. $A^{1,0} + A^{0,1} = I + \xi A \iff A = \xi^{-1}(A^{1,0} + A^{0,1} - I)$.
2. $A^{i,0} = (A^{1,0})^i$, $\forall i \geq 1$, and $A^{0,j} = (A^{0,1})^j$, $\forall j \geq 1$.

The 2-D δ-model's characteristic equation and transfer function, and their relationships to those of the corresponding q-model are as follows:

$$\begin{aligned} \det[I_c - A] &= \det[c_h I_{n_h} \oplus c_v I_{n_v} - A] = \frac{1}{\det[\xi]} \det[I_z - \hat{A}]|_{\mathbf{z}\to\mathbf{c}} \in \Re[c_h]_{n_h}[c_v]_{n_v}; \\ H(c_h, c_v) &= C(I_c - A)^{-1}B + D = \hat{H}(z_h, z_v)|_{\mathbf{z}\to\mathbf{c}} \in \Re(c_h)_{n_h}(c_v)_{n_v}, \end{aligned} \tag{3.10}$$

where

$$\begin{aligned} c_h &= \frac{z_h - 1}{\tau_h} \iff z_h = 1 + \tau_h c_h; \\ c_v &= \frac{z_v - 1}{\tau_v} \iff z_v = 1 + \tau_v c_v. \end{aligned} \tag{3.11}$$

As for the q-model, it is easy to show that, 2-D equivalent transformations of the type

$$\begin{aligned} \begin{bmatrix} \tilde{\mathbf{x}}^h(i,j) \\ \tilde{\mathbf{x}}^v(i,j) \end{bmatrix} &= \begin{bmatrix} T^{(1)} & \mathbf{0} \\ \mathbf{0} & T^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} \\ &\doteq [T] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}, \end{aligned} \tag{3.12}$$

where $T^{(1)} \in \Re^{n_h \times n_h}$ and $T^{(4)} \in \Re^{n_v \times n_v}$ are nonsingular, yield the equivalent 2-D state-space realization $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ where

$$\tilde{A} = TAT^{-1}, \quad \tilde{B} = TB, \quad \tilde{C} = CT^{-1}, \quad \text{and} \quad \tilde{D} = D. \tag{3.13}$$

The transfer function of the realization $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ is the same as that for the realization $\{A, B, C, D\}$.

We will also assume that

$$\det[I_c - A] \neq 0, \ \forall (c_h, c_v) \in \overline{\mathcal{U}}_\delta^2. \tag{3.14}$$

Due to (2.4) and (3.14), assuming no nonessential singularities of the second kind on $\mathcal{T}_\delta^2$, this implies BIBO stability of the 2-D δ-model (see Premaratne and Boujarwah 1994, and references therein).

3.3. Gramians

In the 2-D q-operator case, the reachability and observability gramians are typically taken

to be natural extensions of the integral expressions of their 1-D counterparts (see Premaratne, et. al. 1990, and references therein). In order to adopt a similar approach for the δ-operator case, we first need to investigate the 1-D gramians for the δ-operator case as defined in Middleton and Goodwin (1990).

1-D case. We quote the relevant definitions from Middleton and Goodwin (1990), p. 194 and 200:

DEFINITION 3.1. (Middleton and Goodwin 1990). Consider the 1-D stable δ-system $\{A, B, C, D\}$. The reachability gramian P and observability gramian Q are defined such that they satisfy the following Lyapunov equations:

$$AP + PA^* + \xi \cdot APA^t = -BB^*;$$
$$A^TQ + QA + \xi \cdot A^TQA = -C^*C.$$

We now provide the integral representations of P and Q:

LEMMA 3.1. Consider the 1-D stable δ-system $\{A, B, C, D\}$ with gramians P and Q. Let $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ with gramians $\hat{P}$ and $\hat{Q}$ be the analogous 1-D stable q-system. Then

$$P = \frac{1}{2\pi j} \oint_{T_q} F(c)F^*(c) \frac{dc}{1 + \xi c};$$
$$Q = \frac{1}{2\pi j} \oint_{T_q} G^*(c)G(c) \frac{dc}{1 + \xi c}.$$

Moreover

$$P = \frac{1}{\xi}\hat{P} \iff \hat{P} = \xi P;$$
$$Q = \xi\hat{Q} \iff \hat{Q} = \frac{1}{\xi}Q.$$

Proof. Note that, $\hat{A} = I_n + \xi A$, $\hat{B} = \xi B$, $\hat{C} = C$, and $\hat{D} = D$ (Middleton and Goodwin 1990). Substitute these relationships in the Lyapunov equation for P in Definition 3.1 to get

$$A^*P\hat{A}^* - P = -\frac{1}{\xi}\hat{B}\hat{B}^*.$$

Noting that $\hat{P}$ must satisfy

$$A^*\hat{P}\hat{A}^* - \hat{P} = -\hat{B}\hat{B}^*,$$

we have $P = \hat{P}/\xi$. Moreover, the integral expression for $\hat{P}$ is

$$\hat{P} = \frac{1}{2\pi j} \oint_{T_q} \hat{F}(z)\hat{F}^*(z) \frac{dz}{z},$$

where $\hat{F}(z) \doteq (zI_n - \hat{A})^{-1}\hat{B}$ (Lu, et. al. 1986). The claim regarding P now follows. The proof regarding Q follows in a similar manner. ■

2-D case. With Lemma 3.1 in mind, we now present the following

DEFINITION 3.2. Consider the 2-D stable δ-system $\{A, B, C, D\}$. The reachability gramian P and observability gramian Q are defined as

$$P \doteq \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} F(c_h, c_v) F^*(c_h, c_v) \frac{dc_h}{1 + \tau_h c_h} \frac{dc_v}{1 + \tau_v c_v};$$

$$Q \doteq \begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} G^*(c_h, c_v) G(c_h, c_v) \frac{dc_h}{1 + \tau_h c_h} \frac{dc_v}{1 + \tau_v c_v},$$

where

$$F(c_h, c_v) \doteq (I_c - A)^{-1} B = \begin{bmatrix} \mathbf{f}_1^* \\ \mathbf{f}_2^* \\ \vdots \\ \mathbf{f}_n^* \end{bmatrix} \in \Re^{n \times p}(c_h)_{n_h}(c_v)_{n_v};$$

$$G(c_h, c_v) \doteq C(I_c - A)^{-1} = [\, \mathbf{g}_1 \quad \mathbf{g}_2 \quad \cdots \quad \mathbf{g}_n \,] \in \Re^{q \times n}(c_h)_{n_h}(c_v)_{n_v}.$$

Remarks.

1. Note that, $\mathbf{f}_i(c_h, c_v) \in \Im^p$, $\forall i = 1, \ldots, n$, and $\mathbf{g}_j(c_h, c_v) \in \Im^q$, $\forall j = 1, \ldots, n$.
2. To eventually compare the performance of the δ-model and its corresponding q-model, the following relationships will be useful:

$$(I_c - A)^{-1}|_{\mathbf{c}\to\mathbf{z}} = (I_z - \hat{A})^{-1}\xi;$$

$$F(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \hat{F}(z_h, z_v) \Longleftrightarrow \mathbf{f}_j(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \hat{\mathbf{f}}_j, \quad \text{for} \quad j = 1, \ldots, n;$$

$$G(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \hat{G}(z_h, z_v) \cdot \xi \Longleftrightarrow \mathbf{g}_i(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \begin{cases} \tau_h \hat{\mathbf{g}}_i, & \text{for } i = 1, \ldots, n_h; \\ \tau_v \hat{\mathbf{g}}_i, & \text{for } i = n_h + 1, \ldots, n. \end{cases} \tag{3.15}$$

3. The above definition is completely analogous to the 1-D and 2-D q-operator cases. In the latter case, these gramians have been extremely useful in, and hence, have been extensively used for, investigating coefficient sensitivity, roundoff noise propagation, model reduction, etc. For instance, see Lin, et. al. (1986), (1987), Lu, et. al. (1986), Premaratne, et. al. (1990), and references therein.

LEMMA 3.2. Consider the 2-D stable δ-system $\{A, B, C, D\}$ with gramians P and Q. Let $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ with gramians $\hat{P}$ and $\hat{Q}$ be the analogous 2-D stable q-system. Then

$$P = \frac{1}{\tau_h \tau_v} \hat{P} \iff \hat{P} = \tau_h \tau_v P;$$

$$Q = \frac{1}{\tau_h \tau_v} \xi \hat{Q} \xi \iff \hat{Q} = \tau_h \tau_v \xi^{-1} Q \xi^{-1}.$$

Proof. Consider the integral expression for P in Definition 3.2. With the variable change $\mathbf{c} \rightarrow \mathbf{z}$ and (3.15), we get

$$P = \frac{1}{\tau_h \tau_v} \frac{1}{(2\pi j)^2} \oint \oint_{T_q^2} \hat{F}(z_h, z_v) \hat{F}^*(z_h, z_v) \frac{dz_h}{z_h} \frac{dz_v}{z_v}.$$

However (Premaratne, et. al. 1990),

$$\hat{P} = \frac{1}{(2\pi j)^2} \oint \oint_{T_q^2} \hat{F}(z_h, z_v) \hat{F}^*(z_h, z_v) \frac{dz_h}{z_h} \frac{dz_v}{z_v}.$$

Hence, the claim regarding P follows. The proof regarding Q is similar. ∎

COROLLARY 3.3. The block matrices of the gramians are related as follows:

$$\begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix} = \frac{1}{\tau_h \tau_v} \begin{bmatrix} \hat{P}^{(1)} & \hat{P}^{(2)} \\ \hat{P}^{(3)} & \hat{P}^{(4)} \end{bmatrix} \iff \begin{bmatrix} \hat{P}^{(1)} & \hat{P}^{(2)} \\ \hat{P}^{(3)} & \hat{P}^{(4)} \end{bmatrix} = \tau_h \tau_v \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix};$$

$$\begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{\tau_h}{\tau_v} \hat{Q}^{(1)} & \hat{Q}^{(2)} \\ \hat{Q}^{(3)} & \frac{\tau_v}{\tau_h} \hat{Q}^{(4)} \end{bmatrix} \iff \begin{bmatrix} \hat{Q}^{(1)} & \hat{Q}^{(2)} \\ \hat{Q}^{(3)} & \hat{Q}^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{\tau_v}{\tau_h} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & \frac{\tau_h}{\tau_v} Q^{(4)} \end{bmatrix}.$$

Proof. This follows directly from Lemma 3.2. ∎

With the above results in mind, we now make some pertinent statements that are in complete analogy with the 2-D q-operator case. These may be easily verified/justified from the corresponding results for the latter (see Premaratne, et. al. 1990, and references therein).

LEMMA 3.4. The gramians may be represented as follows:

$$P = \frac{1}{\tau_h \tau_v} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} M_{i,j} M_{ij}^*;$$

$$Q = \frac{1}{\tau_h \tau_v} \xi \cdot \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A^{i,j^*} C^* C A^{i,j} \cdot \xi,$$

where

$$M_{ij} = \begin{cases} 0, & \text{for } (i,j) = (0,0); \\ A^{i-1,j} \xi \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i,j-1} \xi \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix}, & \text{for } (i,j) > (0,0). \end{cases}$$

LEMMA 3.5. Consider the 2-D stable δ-model $\{A, B, C, D\}$ with gramians P and Q. Let $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ with gramians $\tilde{P}$ and $\tilde{Q}$ be an equivalent system obtained with a nonsingular transformation of the type in (3.12-13). Then,

$$\tilde{P} = TPT^* \quad \text{and} \quad \tilde{Q} = T^{-1^*} Q T^{-1}.$$

Moreover, the eigenvalues of PQ are invariant under such a transformation.

DEFINITION 3.3. The 2-D δ-model $\{A, B, C, D\}$ is said to be *balanced* if its gramians P and Q satisfy

$$P^{(1)} = Q^{(1)} \doteq \Sigma^{(1)} = \text{diag}\{\sigma_1^{(1)}, \sigma_2^{(1)}, \ldots, \sigma_{n_h}^{(1)}\};$$
$$P^{(4)} = Q^{(4)} \doteq \Sigma^{(4)} = \text{diag}\{\sigma_1^{(4)}, \sigma_2^{(4)}, \ldots, \sigma_{n_v}^{(4)}\}.$$

If the principal block diagonal matrices of P and Q are each positive definite, a corresponding balanced realization may be obtained through a simultaneous diagonalization procedure (Laub, et. al. 1987). Regarding this, we have

LEMMA 3.6. Local reachability and observability of the δ-model $\{A, B, C, D\}$ and its corresponding q-model $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ are equivalent. Moreover, when $\{A, B, C, D\}$ is locally reachable and observable, $P^{(1)}$, $P^{(4)}$, $Q^{(1)}$, and $Q^{(4)}$ are each positive definite.

Separable systems. A separable (in denominator) 2-D q-system has the property that $\hat{A}^{(2)} = 0$ (or, equivalently, $\hat{A}^{(3)} = 0$). For such a system, Premaratne, et. al. (1990) has shown that, the off-diagonal submatrices of $\hat{P}$ and $\hat{Q}$ are all zero. Moreover, the diagonal submatrices may be conveniently computed through the solution of two pairs of Lyapunov equations.

From (3.7), it is clear that, a separable 2-D q-system gives rise to a separable 2-D δ-system. Regarding the corresponding gramians, we may state the following

THEOREM 3.7. Consider the separable 2-D δ-system $\{A, B, C, D\}$ with gramians P and Q. Then,

$$P^{(2)} = Q^{(2)} = 0 \quad \text{and} \quad P^{(3)} = Q^{(3)} = 0.$$

Moreover, the diagonal block matrices of P and Q may be computed through the solution of the following two pairs of Lyapunov equations:

$$A^{(1)}P^{(1)} + P^{(1)}A^{(1)^*} + \tau_h A^{(1)}P^{(1)}A^{(1)^*} = -\frac{1}{\tau_v}B^{(1)}B^{(1)^*};$$
$$A^{(1)^*}Q^{(1)} + Q^{(1)}A^{(1)} + \tau_h A^{(1)^*}Q^{(1)}A^{(1)} = -\frac{1}{\tau_v}[\,C^{(1)} \quad R^{(4)}A^{(3)}\,]^*[\,C^{(1)} \quad R^{(4)}A^{(3)}\,];$$
$$A^{(4)}P^{(4)} + P^{(4)}A^{(4)^*} + \tau_v A^{(4)}P^{(4)}A^{(4)^*} = -\frac{1}{\tau_h}[\,B^{(2)} \quad A^{(3)}S^{(1)}\,][\,B^{(2)} \quad A^{(3)}S^{(1)}\,]^*;$$
$$A^{(4)^*}Q^{(4)} + Q^{(4)}A^{(4)} + \tau_v A^{(4)^*}Q^{(4)}A^{(4)} = -\frac{1}{\tau_h}C^{(2)^*}C^{(2)},$$

where $R^{(4)^*}R^{(4)} = \tau_h\tau_v Q^{(4)}$ and $S^{(1)}S^{(1)^*} = \tau_h\tau_v P^{(1)}$.

Proof. The results regarding the off-diagonal submatrices are obsious from Corollary 3.3. Regarding the diagonal submatrices, the claim may be shown using Theorem 3.2.2 of Premaratne, et. al. 1990. For instance, consider the Lypaunov equation

$$\hat{A}^{(1)^*}\hat{Q}^{(1)}A^{(1)} - \hat{Q}^{(1)} = -\hat{C}^{(1)^*}\hat{C}^{(1)} - \hat{A}^{(3)^*}\hat{Q}^{(4)}\hat{A}^{(3)}.$$

Using (3.7) and Corollary 3.3, the second Lyapunov equation in the claim results. The rest follows in a similar manner. ■

IV. Coefficient Sensitivity

Coefficient sensitivity is an important criterion on which one state-space realization may be preferred over another. In practice, effects of coefficient sensitivity appears in the system frequency response. Hence, it is important to study the quantities $\partial H/\partial A$, $\partial H/\partial B$, $\partial H/\partial C$ and $\partial H/\partial D$.

By generalizing a certain sensitivity measure in Tavsanoglu and Thiele (1984), Lutz and Hakimi (1988) have addressed sensitivity minimization of MIMO 1-D continuous-time systems. The 2-D q-operator case appears in Lin, et. al. (1987), and references therein. This work, applicable only to the SISO case, has revealed that realizations possessing minimum coefficient sensitivity are equivalent to balanced (modulo a block orthogonal similarity transformation) realizations (see Premaratne, et. al. 1990, and references therein).

In what follows, we study the coefficient sensitivity properties of the 2-D δ-model introduced in Section III. Both FXP and FLP arithmetic implementations are addressed. We follow a more direct approach approach through the use of Kronecker product formulation and, as a result, the results are applicable to the more general MIMO case. Relationships regarding matrix Kronecker products are taken from the excellent treatise of Brewer (1978) and, for the readers' convenience, where appropriate, the equation numbers of Brewer (1978)—these begin with the letter T—are indicated.

First,

$$\begin{aligned} S_A(c_h, c_v) \doteq \frac{\partial}{\partial A} H(c_h, c_v) &= \frac{\partial}{\partial A}[C(I_c - A)^{-1}B + D] \\ &= [I_n \otimes C][I_n \otimes (I_c - A)^{-1}] \cdot \frac{\partial}{\partial A}[I_c - A] \cdot [I_n \otimes (I_c - A)^{-1}][I_n \otimes B] \\ &\qquad \text{from (T4.3) and (T5.5)} \\ &= [I_n \otimes C(I_c - A)^{-1}] \cdot \overline{U}_{n\times n} \cdot [I_n \otimes (I_c - A)^{-1}B] \\ &\qquad \text{from (T2.4) and (T5.1).} \end{aligned}$$

Hence

$$S_A(c_h, c_v) = [I_n \otimes G] \cdot \overline{U}_{n\times n} \cdot [I_n \otimes F] \in \Im^{nq\times np}. \tag{4.1}$$

Second,

$$\begin{aligned} S_B(c_h, c_v) \doteq \frac{\partial}{\partial B} H(c_h, c_v) &= \frac{\partial}{\partial B}[C(I_c - A)^{-1}B + D] = \frac{\partial}{\partial B}[GB] \\ &= [I_n \otimes G] \cdot \frac{\partial B}{\partial B} \quad \text{from (T4.3).} \end{aligned}$$

Hence

$$S_B(c_h, c_v) = [I_n \otimes G] \cdot \overline{U}_{n\times p} \in \Im^{nq\times p^2}. \tag{4.2}$$

Third,

$$\begin{aligned} S_C(c_h, c_v) &\doteq \frac{\partial}{\partial C} H(c_h, c_v) = \frac{\partial}{\partial C}[C(I_c - A)^{-1}B + D] = \frac{\partial}{\partial C}[CF] \\ &= \frac{\partial C}{\partial C} \cdot [I_n \otimes F] \quad \text{from (T4.3).} \end{aligned}$$

Hence

$$S_C(c_h, c_v) = \overline{U}_{q\times n} \cdot [I_n \otimes F] \in \Im^{q^2\times np}. \tag{4.3}$$

Fourth,

$$\begin{aligned} S_D(c_h, c_v) &\doteq \frac{\partial}{\partial D} H(c_h, c_v) = \frac{\partial}{\partial D}[C(I_c - A)^{-1}B + D] \\ &= \frac{\partial D}{\partial D}. \end{aligned}$$

Hence

$$S_D(c_h, c_v) = \overline{U}_{q\times p} \in \Re^{q^2\times p^2}. \tag{4.4}$$

LEMMA 4.1 The quantities $S_A(c_h, c_v)$, $S_B(c_h, c_v)$, $S_C(c_h, c_v)$, and $S_D(c_h, c_v)$ of the δ-model are given as follows:

$$S_A(c_h, c_v) = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} [\, \mathbf{f}_1^* \quad \mathbf{f}_2^* \quad \cdots \quad \mathbf{f}_n^* \,];$$

$$S_B(c_h, c_v) = \begin{bmatrix} \mathbf{g}_1^{(1)} & \mathbf{g}_1^{(2)} & \cdots & \mathbf{g}_1^{(p)} \\ \mathbf{g}_2^{(1)} & \mathbf{g}_2^{(2)} & \cdots & \mathbf{g}_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_n^{(1)} & \mathbf{g}_n^{(2)} & \cdots & \mathbf{g}_n^{(p)} \end{bmatrix};$$

$$S_C(c_h, c_v) = \begin{bmatrix} \mathbf{f}_1^{(1)^*} & \mathbf{f}_2^{(1)^*} & \cdots & \mathbf{f}_n^{(1)^*} \\ \mathbf{f}_1^{(2)^*} & \mathbf{f}_2^{(2)*} & \cdots & \mathbf{f}_n^{(2)^*} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{f}_1^{(q)^*} & \mathbf{f}_2^{(q)^*} & \cdots & \mathbf{f}_n^{(q)^*} \end{bmatrix};$$

$$S_D(c_h, c_v) = \begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,p} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ E_{q,1} & E_{q,2} & \cdots & E_{q,p} \end{bmatrix}.$$

Here, $\mathbf{f}_i^{(j)^*}$ denotes a $(q \times p)$ null matrix except its j-th row which is $\mathbf{f}_i^*$, $\mathbf{g}_i^{(j)}$ denotes a $(q \times p)$ null matrix except its j-th column which is $\mathbf{g}_i$, and $E_{i,j}$ are $(n \times p)$ elementary matrices.

Proof. The relationship for S_D follows immediately from (4.4). To show the remainder, note that

$$[I_n \otimes F] = \begin{bmatrix} F & 0 & \cdots & 0 \\ 0 & F & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & F \end{bmatrix} \in \Im^{n^2 \times np} \text{ and } [I_n \otimes G] = \begin{bmatrix} G & 0 & \cdots & 0 \\ 0 & G & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & G \end{bmatrix} \in \Im^{nq \times n^2}.$$

Here, $[I_n \otimes F]$ and $[I_n \otimes G]$ each has $(n \times n)$ blocks. The claim now follows through simple yet tedious algebraic manipulations. ∎

COROLLARY 4.2. The quantities $S_A(c_h, c_v)$, $S_B(c_h, c_v)$, $S_C(c_h, c_v)$, and $S_D(c_h, c_v)$ of the δ-model and the quantities $\hat{S}_{\hat{A}}(z_h, z_v)$, $\hat{S}_{\hat{B}}(z_h, z_v)$, $\hat{S}_{\hat{C}}(z_h, z_v)$, and $\hat{S}_{\hat{D}}(z_h, z_v)$ of the corresponding q-model are related through the following:

$$S_A(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \Xi \hat{S}_{\hat{A}}(z_h, z_v); \qquad S_B(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \Xi \hat{S}_{\hat{B}}(z_h, z_v);$$
$$S_C(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \hat{S}_{\hat{C}}(z_h, z_v); \qquad S_D(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \hat{S}_{\hat{D}}(z_h, z_v),$$

where $\Xi \doteq [\tau_h I_{n_h q} \oplus \tau_v I_{n_v q}]$.

Proof. This is immediate when (3.15) is applied to Lemma 4.1. ∎

To proceed further, we utilize the following

DEFINITION 4.1. Let $H(c_h, c_v)$ be a bivariate matrix-valued function that is analytic on $\mathcal{T}_\delta^2$. Then,

$$\|H(c_h, c_v)\|_p \doteq \left[\frac{1}{(2\pi\jmath)^2} \oint \oint_{T_q^2} \|H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\|_F^p \frac{dz_h}{z_h} \frac{dz_v}{z_v} \right]^{\frac{1}{p}}$$
$$= \left[\frac{1}{(2\pi)^2} \int_{\omega_h=0}^{2\pi} \int_{\omega_v=0}^{2\pi} \|H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\|_F^p \, d\omega_h \, d\omega_v \right]^{\frac{1}{p}}.$$

Remark. This matrix norm is extensively utilized in work related to coefficient sensitivity (see Lin, et. al. 1987, and references therein) due mainly to the fact that it leads to tractable results. This, and our desire to make a comparison with the corresponding q-model, are the primary reasons for its use here.

FXP Arithmetic Case

Assuming the actual implementation is carried out using FXP arithmetic, we now define an

absolute sensitivity measure that takes into account the variations in the transfer function $H(c_h, c_v)$ with respect to perturbations in A, B, C, and D as follows:

$$\begin{aligned} M_{\text{FXP}} \doteq & \left(\sum\sum \frac{\partial H}{\partial a_{ij}}\right)^2 + \frac{1}{p}\left(\sum\sum \frac{\partial H}{\partial b_{ij}}\right)^2 + \frac{1}{q}\left(\sum\sum \frac{\partial H}{\partial c_{ij}}\right)^2 \\ & + \frac{1}{pq}\left(\sum\sum \frac{\partial H}{\partial d_{ij}}\right)^2 \\ = & \|S_A\|_1^2 + \frac{1}{p}\|S_B\|_2^2 + \frac{1}{q}\|S_C\|_2^2 + \frac{1}{pq}\|S_D\|_2^2. \end{aligned} \tag{4.5}$$

Remarks.

1. The use of different norms is for mathematical feasibility and tractability, and is typical in coefficient sensitivity studies (Lin, et. al. 1987, Li and Gevers 1990). Given a δ-model $\{A, B, C, D\}$, the objective is to characterize those realizations belonging to the class $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\} \equiv \{TAT^{-1}, TB, CT^{-1}, D\}$, where T is a nonsingular equivalent transformation of the type in (3.12), that minimize M_{FXP}.
2. The weights associated with each term in (4.5) may be thought of as *averaging factors*. The ensuing measure then may be thought of as an *average sensitivity per input/output.*
3. In a δ-operator implementation, due to the necessity of performing the computation in (3.6), coefficient sensitivity will be affected by perturbation of τ_h and τ_v as well. Hence, M_{FXP} must be modified to contain terms of the nature $\|S_{\tau_h}\|_2$ and $\|S_{\tau_v}\|_2$. However, the selection of τ_h and τ_v may be done somewhat arbitrarily so that they possess exact binary FXP representations. If so, the corresponding sensitivity terms may be neglected. In what follows, we therefore assume that τ_h and τ_v have been selected as above.

Now, we are in a position to attempt to obtain an expression for M_{FXP} as follows:

$$\begin{aligned} \|S_A\|_1^2 = & \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2}\left\|\begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix}\begin{bmatrix} \mathbf{f}_1^* & \cdots & \mathbf{f}_n^* \end{bmatrix}|_{\mathbf{c}\to\mathbf{z}}\right\|_F \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right]^2 \\ \leq & \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2}\left\|\begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix}|_{\mathbf{c}\to\mathbf{z}}\right\|_F^2 \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right] \\ & \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2}\|\begin{bmatrix} \mathbf{f}_1^* & \cdots & \mathbf{f}_n^* \end{bmatrix}|_{\mathbf{c}\to\mathbf{z}}\|_F^2 \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right] \\ = & \text{trace}\left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2} G^*(c_h, c_v)G(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\frac{dz_h}{z_h}\frac{dz_v}{z_v}\right] \end{aligned}$$

$$\cdot \operatorname{trace}\left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{\mathcal{T}_q^2} F(c_h,c_v)F^*(c_h,c_v)|_{\mathbf{c}\to\mathbf{z}}\frac{dz_h}{z_h}\,\frac{dz_v}{z_v}\right].$$

To get the first inequality, we have used the mutual consistency of Fröbenius norm, that is, $\|AB\|_F \le \|A\|_F \cdot \|B\|_F$, and Cauchy-Schwarz inequality; the last equality follows due to $\|A\|_F^2 = \operatorname{trace}[A^*A]$ (Golub and Van Loan 1983). Hence, using (3.15),

$$\|S_A\|_1^2 \le \operatorname{trace}[\hat{P}]\cdot\operatorname{trace}[\xi\hat{Q}\xi] = (\tau_h\tau_v)^2\cdot\operatorname{trace}[P]\cdot\operatorname{trace}[Q]. \tag{4.6a}$$

Next,

$$\begin{aligned}
\|S_B\|_2^2 &= \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{\mathcal{T}_q^2}\left\|\begin{bmatrix} \mathrm{g}_1^{(1)} & \cdots & \mathrm{g}_1^{(p)} \\ \vdots & \ddots & \vdots \\ \mathrm{g}_n^{(1)} & \cdots & \mathrm{g}_n^{(p)}\end{bmatrix}|_{\mathbf{c}\to\mathbf{z}}\right\|_F^2\frac{dz_h}{z_h}\,\frac{dz_v}{z_v}\right] \\
&= \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{\mathcal{T}_q^2} p\|G(c_h,c_v)|_{\mathbf{c}\to\mathbf{z}}\|_F^2\frac{dz_h}{z_h}\,\frac{dz_v}{z_v}\right] \\
&= p\cdot\operatorname{trace}\left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{\mathcal{T}_q^2} G^*(c_h,c_v)G(c_h,c_v)|_{\mathbf{c}\to\mathbf{z}}\frac{dz_h}{z_h}\,\frac{dz_v}{z_v}\right]
\end{aligned}$$

Hence,

$$\|S_B\|_2^2 = p\cdot\operatorname{trace}[\xi\hat{Q}\xi] = p\tau_h\tau_v\cdot\operatorname{trace}[Q]. \tag{4.6b}$$

Similarly, we get

$$\|S_c\|_2^2 = q\cdot\operatorname{trace}[\hat{P}] = q\tau_h\tau_v\cdot\operatorname{trace}[P], \tag{4.6c}$$

and

$$\|S_D\|_2^2 = pq. \tag{4.6d}$$

Remark. In a manner that parallels the above, corresponding to (4.6), for the q-system counterpart, we have

$$\begin{aligned}
\|\hat{S}_{\hat{A}}\|_1^2 &\le \operatorname{trace}[\hat{P}]\cdot\operatorname{trace}[\hat{Q}] \\
&= (\tau_h\tau_v)^2\cdot\operatorname{trace}[P]\cdot\operatorname{trace}[\xi^{-1}Q\xi^{-1}]; \qquad (4.7.a)\\
\|\hat{S}_{\hat{B}}\|_2^2 &= p\cdot\operatorname{trace}[\hat{Q}] \\
= p\tau_h\tau_v\cdot\operatorname{trace}[\xi^{-1}Q\xi^{-1}]&; \qquad (4.7b)\\
\|\hat{S}_{\hat{C}}\|_2^2 &= q\cdot\operatorname{trace}[\hat{P}] \\
&= q\tau_h\tau_v\cdot\operatorname{trace}[P]; \qquad (4.7.c)\\
\|\hat{S}_{\hat{D}}\|_2^2 &= pq. \qquad (4.7d)
\end{aligned}$$

Combining (4.5) with (4.6), we get the following upper bound for M_{FXP}:

$$\begin{aligned} M_{\text{FXP}} \leq \overline{M}_{\text{FXP}} &\doteq (\text{trace}[\hat{P}] + 1)(\text{trace}[\xi\hat{Q}\xi] + 1) \\ &= (\tau_h\tau_v \cdot \text{trace}[P] + 1)(\tau_h\tau_v \cdot \text{trace}[Q] + 1). \end{aligned} \tag{4.8}$$

Due to difficulties associated with minimization of M_{FXP}, it is customary to perform a minimization of $\overline{M}_{\text{FXP}}$. Hence, one attempts to characterize those realization $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ that are 'bound optimal' with respect to the sensitivity measure M_{FXP}.

For reasons of brevity, we do not attempt to perform this since the procedure is exactly analogous to the 2-D q-operator case (see Lin, et. al. 1987, and references therein). For instance, one may show that any realization that is balanced modulo an orthogonal non-singular transformation is bound optimal with regards to the sensitivity measure defined in (4.5).

Remark. In a manner that parallels the above, corresponding to (4.8), for the q-system counterpart, we have

$$\begin{aligned} \hat{M}_{\text{FXP}} \leq \overline{\hat{M}}_{\text{FXP}} &\doteq (\text{trace}[\hat{P}] + 1)(\text{trace}[\hat{Q}] + 1) \\ &= (\tau_h\tau_v \cdot \text{trace}[P] + 1)(\tau_h\tau_v \cdot \text{trace}[\xi^{-1}Q\xi^{-1}] + 1). \end{aligned} \tag{4.9}$$

However, it is instructive to note that, compared to a q-operator implementation, its δ-model implementation will always yield a smaller $\overline{M}_{\text{FXP}}$ whenever

$$\begin{aligned} &\text{trace}[\hat{Q}] > \text{trace}[\xi\hat{Q}\xi] \\ &\quad \Leftrightarrow (1 - \tau_h^2) \cdot \text{trace}[\hat{Q}^{(1)}] + (1 - \tau_v^2) \cdot \text{trace}[\hat{Q}^{(4)}] > 0. \end{aligned} \tag{4.10}$$

Note that, with the local reachability and observability assumption of $\{A, B, C, D\}$ (and hence $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$), positive definiteness of $Q^{(1)}$ and $Q^{(4)}$ (and hence $\hat{Q}^{(1)}$ and $\hat{Q}^{(4)}$) are guaranteed. This implies strict positivity of $\text{trace}[Q^{(1)}]$ and $\text{trace}[Q^{(4)}]$ (and hence $\text{trace}[\hat{Q}^{(1)}]$ and $\text{trace}[\hat{Q}^{(4)}]$). Thus, (4.10) is satisfied, that is, the δ-operator implementation is superior with regards to coefficient sensitivity, whenever

$$\tau_h < 1 \quad \text{and} \quad \tau_v < 1. \tag{4.11}$$

FLP Arithmetic Case

If the actual implementation is carried out using FLP arithmetic, the absolute stability

measure in (4.5) may not be the most appropriate. This is because, in FLP implementations, possible perturbation of a particular coefficient will be in fact approximately proportional to its nominal value. Due to this, Li and Gevers (1990), in addressing the 1-D δ-system coefficient sensitivity, utilizes a certain *relative* absolute measure. In the same spirit, a more suitable sensitivity measure for the FLP case will be

$$\begin{aligned} M_{\mathrm{FLP}} \doteq & \left(\sum\sum a_{ij}\frac{\partial H}{\partial a_{ij}}\right)^2 + \frac{1}{p}\left(\sum\sum b_{ij}\frac{\partial H}{\partial b_{ij}}\right)^2 + \frac{1}{q}\left(\sum\sum c_{ij}\frac{\partial H}{\partial c_{ij}}\right)^2 \\ & + \frac{1}{pq}\left(\sum\sum d_{ij}\frac{\partial H}{\partial d_{ij}}\right)^2 \\ = & \|\tilde{S}_A\|_1^2 + \frac{1}{p}\|\tilde{S}_B\|_2^2 + \frac{1}{q}\|\tilde{S}_C\|_2^2 + \frac{1}{pq}\|\tilde{S}_D\|_2^2, \end{aligned} \tag{4.12}$$

where $\tilde{S}_A = \sum\sum a_{ij}\partial H/\partial a_{ij}$, etc. Using Definition 4.1, one may now show that

$$\begin{aligned} \|\tilde{S}_A\|_p &\le \|A\|_F \cdot \|S_A\|_p; \\ \|\tilde{S}_B\|_p &\le \|B\|_F \cdot \|S_B\|_p; \\ \|\tilde{S}_C\|_p &\le \|C\|_F \cdot \|S_C\|_p; \\ \|\tilde{S}_D\|_p &\le \|D\|_F \cdot \|S_D\|_p. \end{aligned} \tag{4.13}$$

Hence, substituting from (3.7), we get

$$M_{rmFLP} \le \|\xi^{-1}(\hat{A}-I)\|_F^2 \cdot \|\xi\hat{S}_{\hat{A}}\|_1^2 + \frac{1}{p}\|\xi^{-1}\hat{B}\|_F^2 \cdot \|\xi\hat{S}_{\hat{B}}\|_2^2 + \frac{1}{q}\|\hat{C}\|_F^2 \cdot \|\hat{S}_{\hat{C}}\|_2^2 + \frac{1}{pq}\|\hat{D}\|_F^2 \cdot \|\hat{S}_{\hat{D}}\|_2^2. \tag{4.14}$$

To proceed farther, let us assume $\tau_h = \tau_v = \tau$ for convenience. Then, we get

$$\begin{aligned} \|\xi^{-1}(\hat{A}-I)\|_F^2 &= \frac{1}{\tau^2}\|\hat{A}-I\|_F^2; \\ \|\xi^{-1}\hat{B}\|_F^2 &= \frac{1}{\tau^2}\|\hat{B}\|_F^2. \end{aligned} \tag{4.15}$$

Combining (4.14) with (4.15), we get the following upper bound for M_{FLP}:

$$M_{\mathrm{FLP}} \le \overline{M}_{\mathrm{FLP}} \doteq \|\hat{A}-I\|_F^2 \cdot \mathrm{trace}[\hat{P}]\mathrm{trace}[\hat{Q}] + \|\hat{B}\|_F^2 \cdot \mathrm{trace}[\hat{Q}] + \|\hat{C}\|_F^2 \cdot \mathrm{trace}[\hat{P}] + \|\hat{D}\|_F^2. \tag{4.16}$$

Again, we perform a minimization of $\overline{M}_{\mathrm{FLP}}$.

Remark. In a manner that parallels the above, corresponding to (4.16), for the q-system counterpart, we have

$$\hat{M}_{\mathrm{FLP}} \le \overline{\hat{M}}_{\mathrm{FLP}} \doteq \|\hat{A}\|_F^2 \cdot \mathrm{trace}[\hat{P}]\mathrm{trace}[\hat{Q}] + \|\hat{B}\|_F^2 \cdot \mathrm{trace}[\hat{Q}] + \|\hat{C}\|_F^2 \cdot \mathrm{trace}[\hat{P}] + \|\hat{D}\|_F^2. \tag{4.17}$$

Hence, compared to a q-operator implementation, its δ-model implementation will yield a smaller $\overline{M}_{\text{FLP}}$ whenever

$$\|\hat{A} - I\|_F^2 < \|\hat{A}\|_F^2. \tag{4.18}$$

Clearly,

$$|\lambda_i[\hat{A}] - 1| < |\lambda_i[\hat{A}]|, \; \forall i = 1, \ldots, n \implies \|\hat{A} - I\|_F^2 < \|\hat{A}\|_F^2, \tag{4.19}$$

where $\lambda_i[\hat{A}]$ denotes the i-th eigenvalue of $\hat{A}$.

Remark. Li and Gevers (1990) refers to the above region (where the eigenvalues of $\hat{A}$ should lie) as the *Middleton-Goodwin (MG) region.* They show that, for the 1-D case, frequency response of a δ-system will be less sensitive to coefficient perturbations if the system eigenvalues lie within the MG region.

V. Example

To illustrate the notions presented previously, let us consider the following stable 2h-2v 2-D digital filter in its Roesser model $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$, where

$$\hat{A} = \begin{bmatrix} 1.8890 & -0.9122 & -1.0000 & 0 \\ 1.0000 & 0 & 0 & 0 \\ 0.0277 & -0.0258 & 1.8890 & 1.0000 \\ -0.0258 & 0.0243 & -0.9122 & 0 \end{bmatrix}; \quad \hat{B} = \begin{bmatrix} 0.1095 \\ 0 \\ -0.0144 \\ 0.0456 \end{bmatrix};$$

$$\hat{C} = [0.1444 \quad -0.0456 \quad -0.1095 \quad 0]; \quad \hat{D} = [0].$$

The magnitude response of this system is shown in Fig. (1). The gramians are computed as

$$\hat{P} = \begin{bmatrix} 21.7931 & 21.3143 & 0.4100 & -0.3848 \\ 21.3143 & 21.7931 & 0.3303 & -0.3083 \\ 0.4100 & 0.3303 & 0.2836 & -0.2589 \\ -0.3848 & -0.3083 & -0.2589 & 0.2434 \end{bmatrix};$$

$$\hat{Q} = \begin{bmatrix} 2.8358 & -2.5893 & 3.9702 & 3.0934 \\ -2.5893 & 2.4168 & -3.7574 & -2.8971 \\ 3.9702 & -3.7574 & 159.8513 & 155.8357 \\ 3.0934 & -2.8971 & 155.8357 & 159.8513 \end{bmatrix}.$$

The similarity transformation

$$\hat{T} = \begin{bmatrix} 2.1813 & -3.6709 & 0 & 0 \\ 1.5578 & -4.1949 & 0 & 0 \\ 0 & 0 & 0.4021 & -0.1581 \\ 0 & 0 & -0.3473 & 0.2247 \end{bmatrix}$$

yields the BL q-system

$$\hat{A}_b = \begin{bmatrix} 0.9664 & 0.1279 & -0.4915 & 0.1932 \\ -0.1611 & 0.9226 & -0.1825 & 0.0718 \\ 0.0463 & 0.0088 & 0.9774 & 0.1747 \\ -0.0103 & -0.0187 & -0.1214 & 0.9116 \end{bmatrix}; \quad \hat{B}_b = \begin{bmatrix} 0.1339 \\ 0.0497 \\ 0.1118 \\ 0.3757 \end{bmatrix};$$

$$\hat{C}_b = [0.2440 \quad -0.3389 \quad -0.0440 \quad 0.0173]; \quad \hat{D}_b = [0].$$

Next, we attempt to obtain the corresponding δ-systems.

FXP Implementation

In FXP, τ_h and τ_v usually determine the range of numbers of the δ-system's s.s. representation. If they are too small, due to (3.7), coefficient values in the δ-system may be too large; if they are too large, the advantages to be gained in terms of coefficient sensitivity may vanish. Hence, these parameters should be carefully selected.

By inspecting the BL q-system, the minimum value for both τ_h and τ_v were selected as 0.5. The corresponding δ-system obtained then has numerical values that are within approximately the same range as for the q-system. Direct conversion of the BL q-system to its correspondiong δ-system gives

$$A = \begin{bmatrix} -0.0672 & 0.2557 & -0.9830 & 0.3864 \\ -0.3222 & -0.1548 & -0.3651 & 0.1435 \\ 0.0926 & 0.0177 & -0.0452 & 0.3494 \\ -0.0207 & -0.0374 & -0.2428 & -0.1768 \end{bmatrix}; \quad B = \begin{bmatrix} 0.2678 \\ 0.0995 \\ 0.2235 \\ 0.7514 \end{bmatrix};$$

$$C = [0.2440 \quad -0.3389 \quad -0.0440 \quad 0.0173]; \quad D = [0].$$

This system is not BL in the sense in Definition 3.3. The BL δ-system is

$$A_b = \begin{bmatrix} -0.0672 & -0.2557 & -0.9830 & -0.3864 \\ 0.3222 & -0.1548 & 0.3651 & 0.1435 \\ 0.0926 & -0.0177 & -0.0452 & -0.3494 \\ 0.0207 & -0.0374 & 0.2428 & -0.1768 \end{bmatrix}; \quad B_b = \begin{bmatrix} 0.1894 \\ -0.0703 \\ 0.1581 \\ -0.5313 \end{bmatrix};$$

$$C_b = [0.3451 \quad 0.4793 \quad -0.0623 \quad -0.0245]; \quad D_b = [0].$$

The BL q- and δ-systems so obtained were implemented in finite precision using FXP. For comparison purposes, the following measures were computed:

$$E_{\max} \doteq \|H(e^{j\omega_i}) - H_{FXP}(e^{j\omega_i})\|_\infty, \quad \text{and}$$
$$E_{\text{sum}} \doteq \sum_{i=0}^{N-1} |H(e^{j\omega_i}) - H_{FXP}(e^{j\omega_i})|^2,$$

where $\omega_i = 2\pi i/N$, N being a sufficiently large positive integer, were computed. Here, $H(e^{j\omega_i})$ is the frequency response in infinite precision and $H_{FXP}(e^{j\omega_i})$ is the frequency response in FXP. The results are shown in Fig. (2).

FLP Implementation

In FLP, a large dynamic range is available. Hence, there is no restriction on the choice of τ_h and τ_v. We select $\tau_h = \tau_v = 1/8$. The resulting BL δ-system is

$$A_b = \begin{bmatrix} -0.2687 & -1.0230 & -3.9321 & -1.5457 \\ 1.2888 & -0.6194 & 1.4602 & 0.5740 \\ 0.3704 & -0.0707 & -0.1807 & -1.3975 \\ 0.0827 & -0.1494 & 0.9710 & -0.7074 \end{bmatrix}; \quad B_b = \begin{bmatrix} 0.3788 \\ -0.1407 \\ 0.3161 \\ -1.0627 \end{bmatrix};$$

$$C_b = [0.6902 \quad 0.9586 \quad -0.1246 \quad -0.0490]; \quad D_b = [0].$$

Again, the BL q- and δ-systems so obtained were implemented in finite precision using FLP. Measures corresponding to $E_{\max}$ and E_{sum} were also computed. The results are shown in Fig. (3).

VII. Conclusion and Final Remarks

In this paper, we have developed a δ-operator based counterpart to the more conventional q-operator based Roesser s.s. model. The motivation for such a development lies in the superior finite wordlength properties exhibited by 1-D δ-operator based discrete-time systems.

The corresponding notions of gramians and BL realization are proposed. For both FXP and FLP implementations, conditions under which the δ-operator formulated system behaves better than its q-operator counterpart are derived. These results indicate that, in most situations, the δ-system, as expected, can be expected to provide superior coefficient sensitivity properties.

In a FXP implementation, however, due to the limited dynamic range available, care must be taken in selecting the 'sampling times' τ_h and τ_v. Of course, in a FLP implementation, such a difficulty does not arise.

This work only addresses the coefficient sensitivity issues. The authors are currently completing work regarding the roundoff error properties of the δ-model developed, where, as in 1-D case, improvements over the corresponding q-model are expected.

We must mention that certain difficulties regarding limit cycles are inherent in δ-systems when FXP arithmetic is used (Premaratne and Bauer 1993). However, this problem is, for all practical purposes, nonexistent in the FLP arithmetic case. Hence, in our opinion, for FLP high-speed applications, the δ-model developed provides an extremely attractive solution that avoids the numrical ill-conditions typically associated with q-systems.

References

[1] J.W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Transactions on Circuits and Systems,* vol. CAS-25, pp. 772-781, Sept. 1978.

[2] E.I. Jury, "Stability of multidimensional systems and other related problems," Chapter 3 in *Multidimensional Systems, Techniques, and Applications,* New York, NY: Marcel Dekkar, 1986.

[3] G.H. Golub and C.F. Van Loan, *Matrix Computations,* Baltimore, MD: John Hopkins University Press, 1983.

[4] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE,* vol. 80, pp. 240-259, 1992.

[5] T. Lin, M. Kawamata, and T. Higuchi, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Transactions on Circuits and Systems,* vol. CAS-33, pp. 724-730, July 1986.

[6] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Transactions on Automatic Control,* vol. AC-32, pp. 115-122, Feb. 1987.

[7] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proceedings of the 1990 IEEE Conference on Decision and Control (CDC'90),* pp. 954-959, Honolulu, HI, Dec. 1990.

[8] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Transactions on Signal Processing,* vol. 41, pp. 629-637, Feb. 1993.

[9] G. Likourezos, "Prolog to 'High-speed digital signal processing and control'," *Proceedings of the IEEE,* vol. 80, pp. 238-239, 1992.

[10] T. Lin, M. Kawamata, and T. Higuchi, "Minimization of sensitivity of 2-D systems and its relation to 2-D balanced realizations," *The Transactions of the IEICE,* vol. E70, pp. 938-944, Oct. 1987; also in *Proceedings of the 1987 IEEE International Symposium on Circuits and Systems (ISCAS'87),* vol. 2, pp. 710-713, Philadeplhia, PA, May 1987.

[11] W.S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital filter structures with minimum roundoff noise," *IEEE Transactions on Circuits and Systems,* vol. CAS-33, pp. 965-973, Oct. 1986.

[12] W.S. Lu and A. Antoniou, *Two-Dimensional Digital Filters,* New York, NY: Marcel Dekker, 1992.

[13] W.S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proceedings of the 1986 IEEE International Symposium on Circuits and Systems (ISCAS'86),* vol. 1, pp. 79-82, 1986.

[14] W.J. Lutz and S.L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Transactions on Circuits and Systems,* vol. 35, pp. 1114-1122, Sept. 1988.

[15] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1990.

[16] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proceedings of the 1994 IEEE International Symposium on Circuits and Systems (ISCAS'94),* London, UK, vol. 2, pp. 461-464, May 1994.

[17] K. Premaratne and A.S. Boujarwah, "Stability determination of two-dimensional delta-operator formulated discrete-time systems," submitted to *Multidimensional Systems and Signal Processing,* 1994.

[18] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Transactions on Circuits and Systems,* vol. CAS-37, pp. 1116-1132, Sept. 1990.

[19] K. Premaratne, R. Salvi, N.R. Habib, and J.P. Le Gall, "Delta-operator formulated discrete-time equivalents of continuous-time systems," *IEEE Transactions on Automatic Control,* vol. 39, pp. 581-585, Mar. 1994.

[20] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Transactions on Automatic Control,* vol. AC-20, pp. 1-10, Feb. 1975.

[21] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Transactions on Circuits and Systems,* vol. CAS-31, pp. 884-888, Oct. 1984.

[21] R. Vijayan, H.V. Poor, J.B. Moore, and G.C. Goodwin, "A Levinson-type algorithm for modeling fast-sampled data," *IEEE Transactions on Automatic Control,* vol. 36, pp. 314-321, Mar. 1991.

Magnitude Response

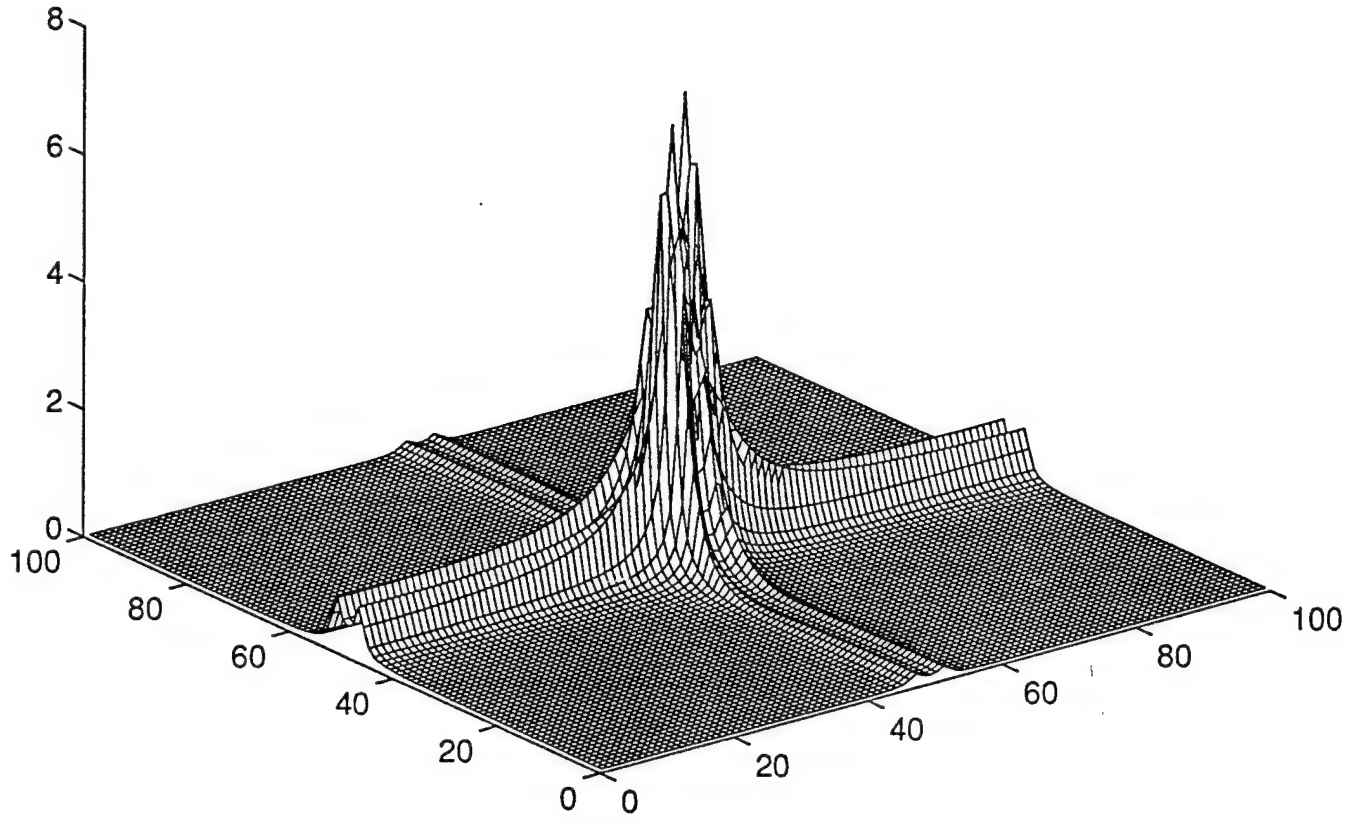
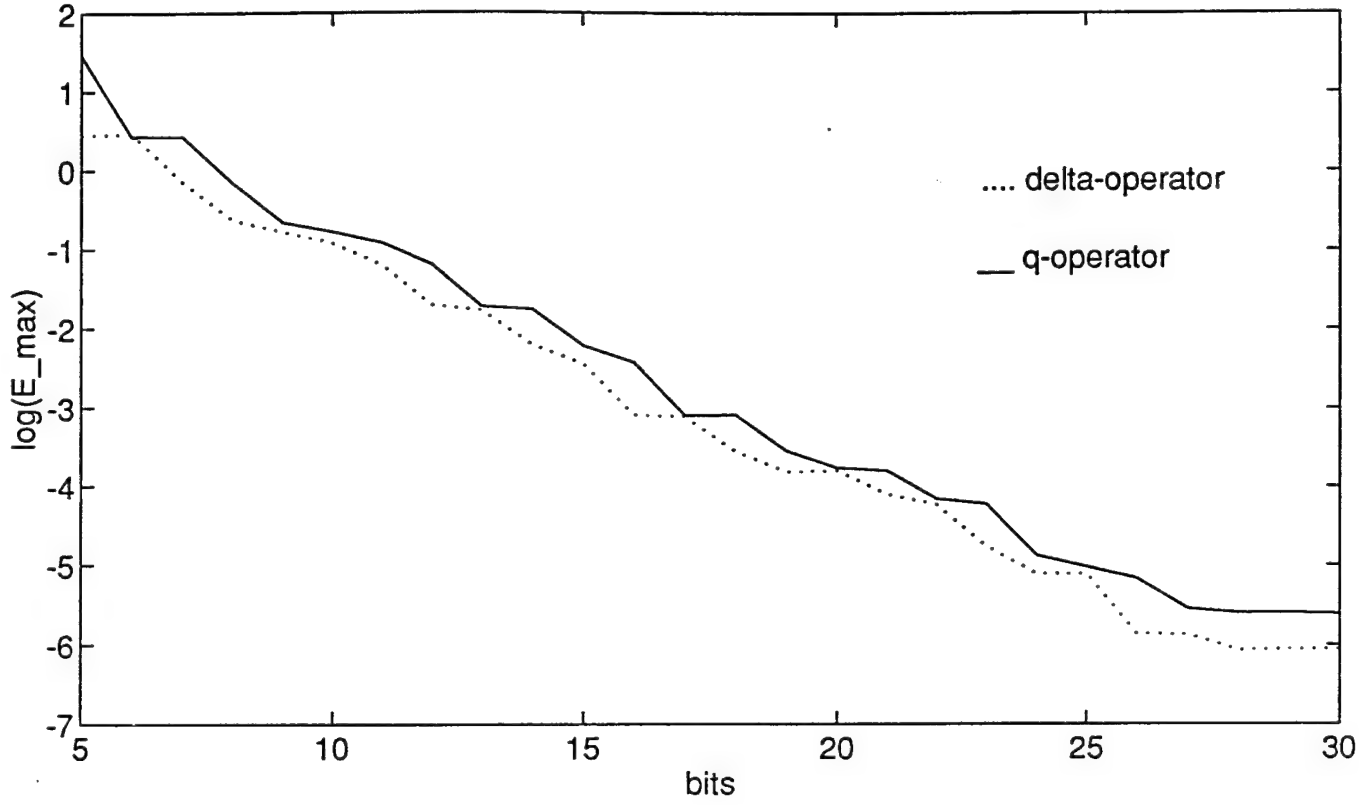


FIG 1


2
1
0
-1
-2
-3
-4
-5
-6
-7
log(E_max)
5
10
15
20
25
30
bits
.... delta-operator
__ q-operator

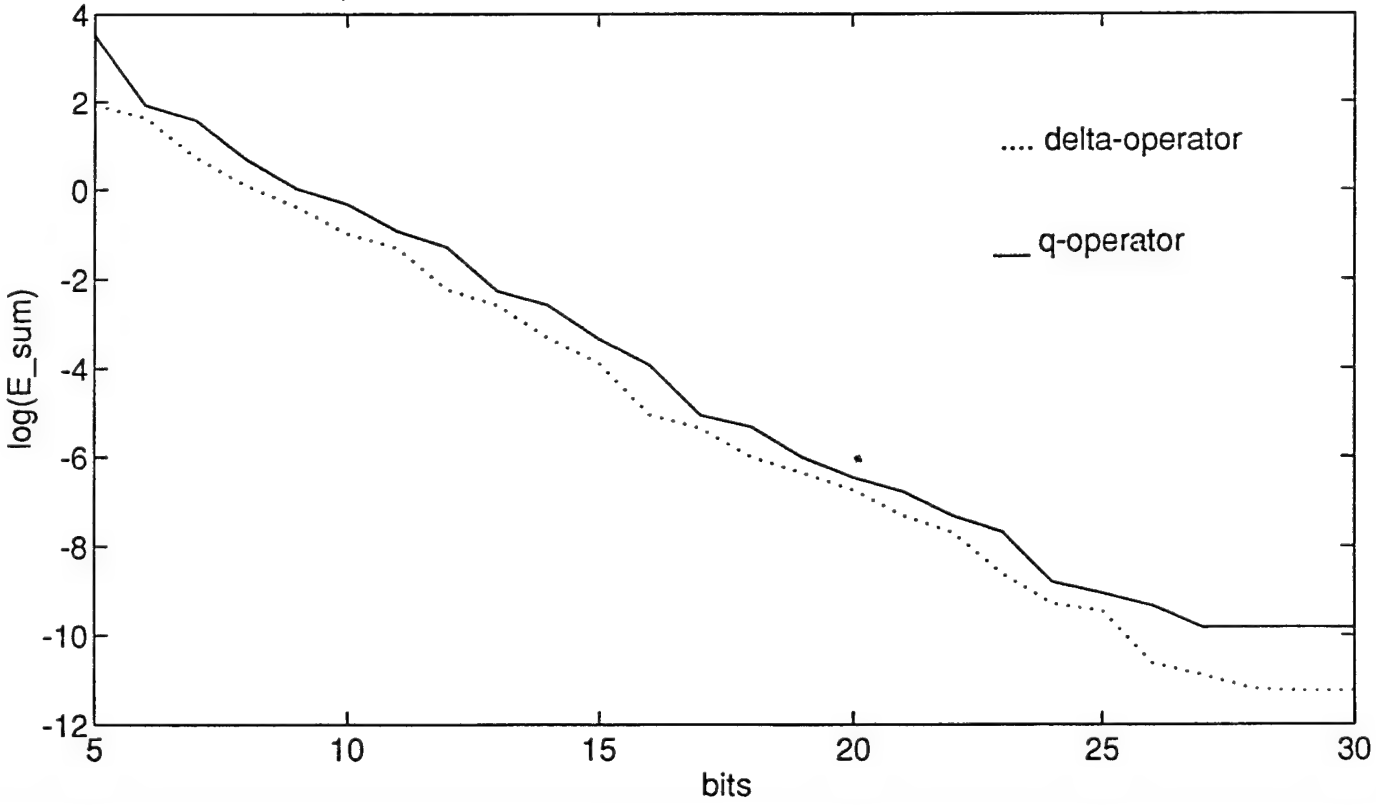
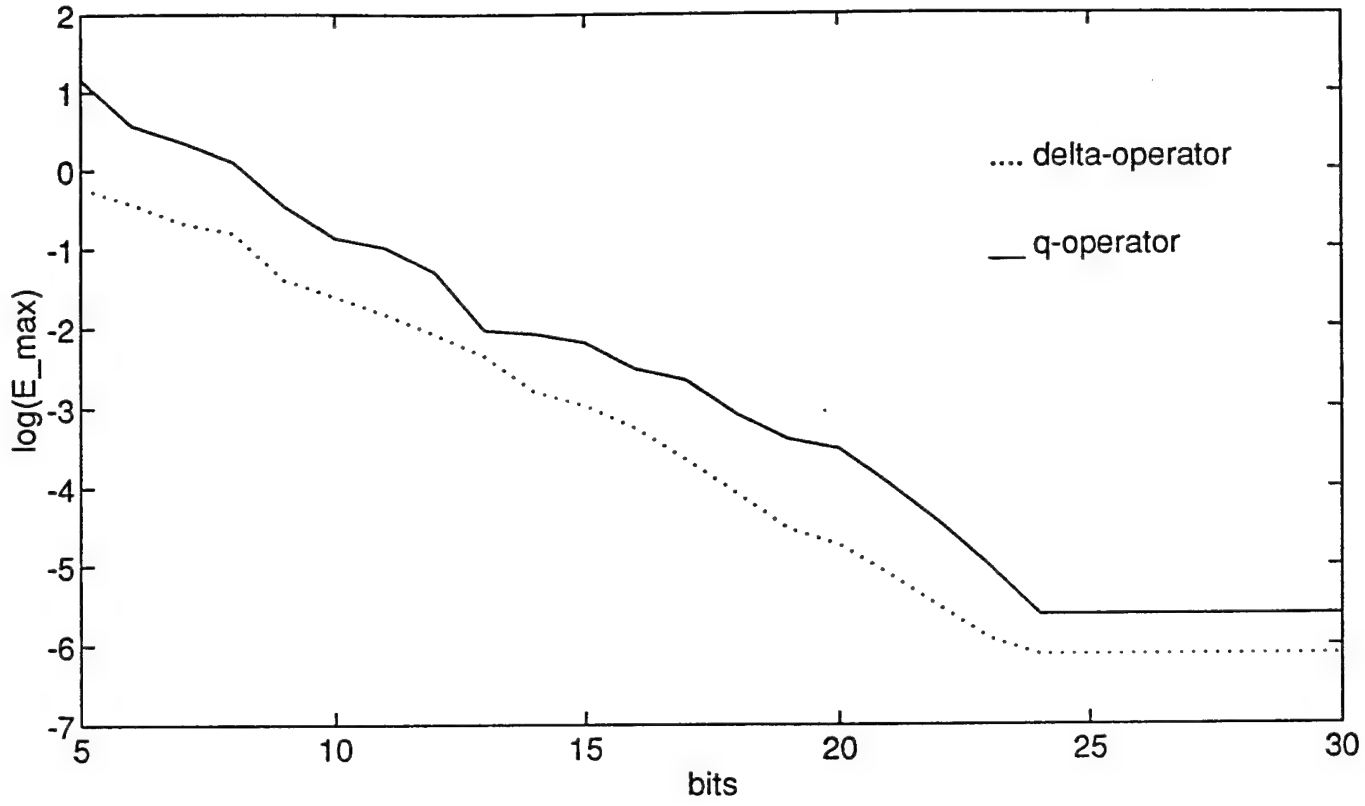

4
2
0
-2
-4
-6
-8
-10
-12
log(E_sum)
5
10
15
20
25
30
bits
.... delta-operator
__ q-operator

FIG 2


log(E_max)
bits
.... delta-operator
__ q-operator
2
1
0
-1
-2
-3
-4
-5
-6
-7
5
10
15
20
25
30

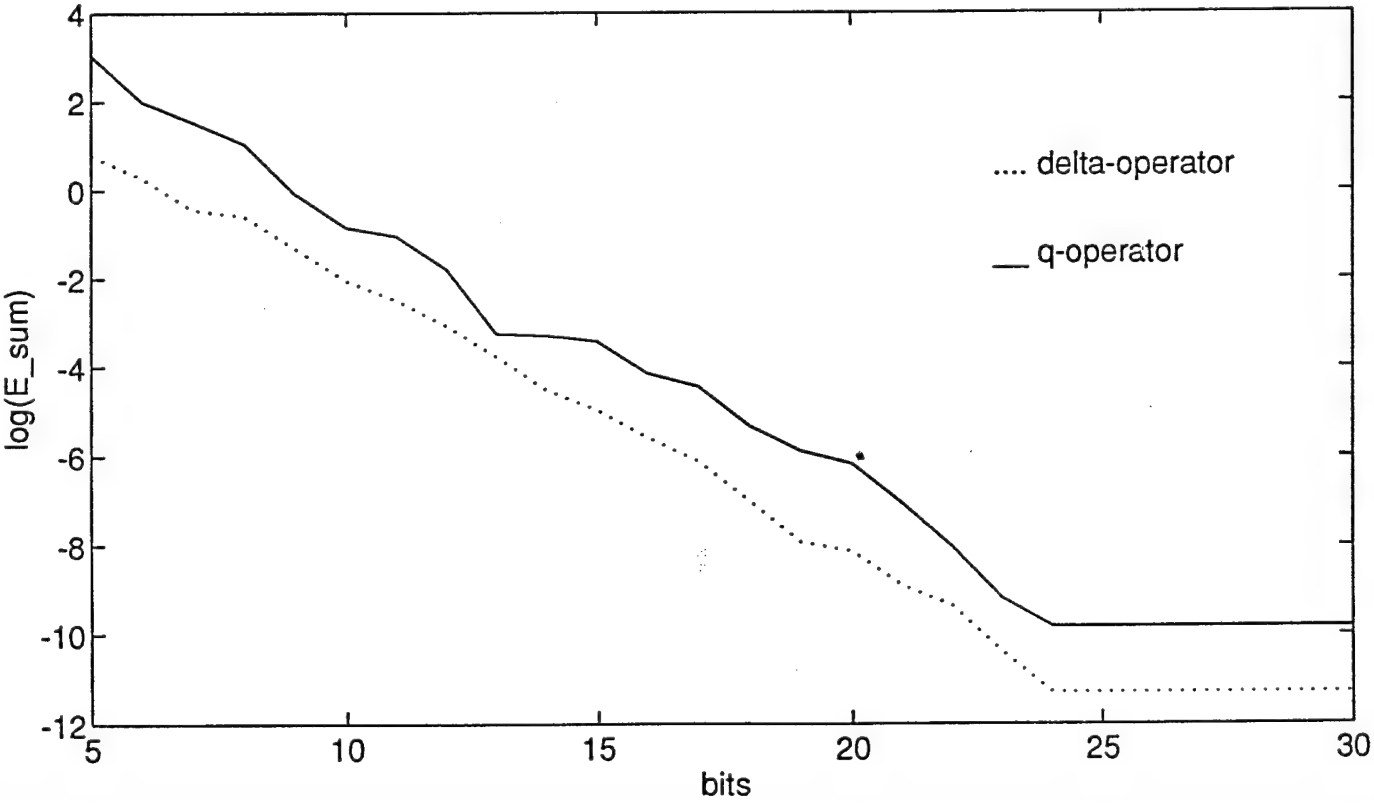

log(E_sum)
bits
.... delta-operator
__ q-operator
4
2
0
-2
-4
-6
-8
-10
-12
5
10
15
20
25
30

FIG 3

An Exhaustive Search Algorithm For Checking Limit Cycle Behavior Of Digital Filters

K. Premaratne, *Senior Member, IEEE*, E.C. Kulasekere [1]
P.H. Bauer[2], *Member, IEEE*, L.J. Leclerc [3]

Abstract :-*The presence of limit cycles that may arise in fixed-point arithmetic implementation of a digital filter can significantly impair its performance. The work in this paper presents an algorithm that can be utilized to determine the presence or the absence of such limit cycles of a given digital filter. The filter is assumed to be in its state-space formulation and hence, performance of the corresponding direct form representation follows as a special case. Moreover, the algorithm is applicable independent of the filter order, type of quantization nonlinearity, and whether the accumulator is single-length or double-length. In developing the algorithm, bounds on the amplitude and period of limit cycles of a given digital filter are obtained. The robustness of the algorithm in terms of limit cycles performance with respect to filter coefficient perturbations is verified. Hence, it may be utilized to obtain regions in the coefficient space where a digital filter of given order is limit cycle free.*

[1]K. Premaratne and E.C. Kulasekere are with the Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124, USA.

[2]P.H. Bauer is with the Department of Electrical Engineering, Laboratory for Signal and Image Analysis (LISA), University of Notre Dame,Notre Dame, IN 46556, USA.

[3]L.J. Leclerc is with the Ericsson Communications, Inc., Ville Mont-Royal, Québec, CANADA.

K.P. and P.H.B. gratefully acknowledge the support received from the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

I Introduction

A digital filter may be realized using either a general purpose digital computer or special purpose digital hardware. In either case, the coefficients and intermediate results of computations must be stored in binary form in registers of finite wordlength. Limit cycle oscillations are a direct result of this limitation, and care must be taken to suppress them while performing a digital filter design.

For the past several years, this in fact has been a research topic of interest, and a significant amount of insight and research results are now available [1]-[10]. In an implementation of a higher order digital filter, as shown in [11], a cascade or parallel form composed of first-order and second-order subfilters is preferable over any direct form realization. Therefore the results are summarized for the second order realizations. Most existing results focus on the effects of signed magnitude rounding and truncation quantization schemes with regard to the existence of limit cycles. Recently, some work addressing the two's complement truncation scheme has also appeared [12]-[14].

This work proposes an algorithm that may be used to check for limit cycles of a given digital filter implemented using fixed-point arithmetic. It possesses a wide scope of applicability: The digital filter to be tested may be of any order; the quantization scheme may be arbitrary, including truncation and rounding schemes corresponding to signed magnitude and two's complement; and the accumulator may be of single- or double-length.

Given a digital filter, we develop bounds on the amplitude and period of possible limit cycles. The algorithm is based on an exhaustive search procedure over all these possibilities. In addition, extending the same procedure to the entire linear stability region, one may utilize it to obtain regions in the filter coefficient space where the given filter is globally asymptotically stable (g.a.s.). For this purpose, the robustness of the algorithm in terms of presence or absence of limit cycles with respect to filter coefficient perturbations is also verified. A similar concept has been used before for checking limit cycle behavior of digital filters implemented in direct form [10], [15]-[18]. The major advantage of the proposed method is that it is applicable for the more general state-space implementations. Of course, the direct form

implementation then follows as a special case.

The paper is organized as follows. Section II contains the nomenclature used throughout the paper. Section III provides bounds on the amplitude and period of limit cycles of a given general digital filter. Section IV discusses the algorithm and its computational aspects. Section V addresses the robustness of the algorithm with respect to perturbations of filter coefficients. Section VI contains some situations where the algorithm developed has been used effectively. Finally, Section VII contains the concluding remarks.

II Nomenclature

The following notation will be used throughout the paper.

$\Re$, $\mathcal{Z}$	Set of reals, set of integers.
$\mathcal{C}$	Set of complex numbers.
$\mathcal{Z}_+$	Nonnegative integers.
$\Re^{m\times n}$, $\mathcal{Z}^{m\times n}$	Set of matrices of size $m \times n$ over the reals and integers.
$\Re(z)_{m\times n}$	Set of matrices of size $m \times n$ over the rational polynomials in the indeterminate $z \in \mathcal{C}$.
$\mathcal{K}[\cdot]$	Cardinality of set $[\cdot]$.
a_{ij}	(i,j)-th element of the matrix $A = \{a_{ij}\}$.
I, $\mathbf{0}$	Identity matrix and null matrix of appropriate sizes.
$\mathbf{x}(k)$	Filter state vector at instant k.
$x_i(k)$	i-th component of the state vector $\mathbf{x}(k)$.
$\|\cdot\|_\infty$	The infinity norm. For $\mathbf{x} = \{x_i\} \in \Re^m$, $\|\mathbf{x}\|_\infty = \max_i \|x_i\|$; for $A = \{a_{ij}\} \in \Re^{m\times n}$, $\|A\|_\infty = \max_i \sum_{j=1}^n \|a_{ij}\|$.
M_i	Upper bound for absolute value of amplitude of $x_i(k)$, $k \in \mathcal{Z}_+$.
$\hat{M}_i$	Largest integer less than or equal to M_i.
$\delta(k)$	Dirac delta function.
$H_{ij}(z)$	(i,j)-th element, that is, the (i,j)-th transfer function, of the MIMO transfer function $H(z)$.
$h_{ij}(k)$	Impulse response of $H_{ij}(z)$.
P_i	i-th pole (accounting for multiplicity) of $H_{ij}(z)$.
K_{ij}	Constant term in the partial fraction expansion of $H_{ij}(z)$.
r_{ij}^k	k-th residue of $H_{ij}(z)$.

q	Quantization step size.
$\mathcal{Q}[\cdot]$	Quantization nonlinearity operator.
ϱ	Normalized quantization error. For instance, for roundoff, $\varrho = 0.5$, and for truncation, $\varrho = 1$.
N	Number of nonlinearities in a realization.
$\mathbf{e}(k)$	Quantization error vector.

III Amplitude and Period Bounds on Limit Cycles.

In general, the quantization nonlinearity satisfies

$$|x - \mathcal{Q}[x]| \leq \varrho \cdot q, \quad \forall x \in \Re, \tag{1}$$

where ϱ is the normalized quantization error. In particular, for roundoff quantization, $\varrho = 0.5$, and for truncation quantization, $\varrho = 1$. Note that, all the filter parameters may be expressed as integer multiples of the quantization step size q. Hence, for convenience, we normalize q to unity for all calculations. The quantization nonlinearity thus becomes an integer valued function, viz.,

$$\mathcal{Q} : \Re \rightarrow \mathcal{Z} \tag{2}$$

In general, for all quantization schemes of interest, $\mathcal{Q}[0] = 0$.

We consider a digital filter of order m in its minimal state-space representation $\{A, B, C, D\}$, that is,

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k) + B \cdot \mathbf{u}(k); \tag{3}$$

$$\mathbf{y}(k) = C \cdot \mathbf{x}(k) + D \cdot \mathbf{u}(k), \tag{4}$$

where $\mathbf{x} \in \Re^m$ is the state, $\mathbf{u}$ is the input, and $\mathbf{y}$ is the output. Also, $A \in \Re^{m \times m}$. For addressing limit cycle performance, we consider the zero input recursive state equation

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k). \tag{5}$$

Unless otherwise stated, we only consider linearly stable filters. Hence, all eigenvalues of A are inside the unit circle in $\mathcal{C}$.

Now, under finite wordlength conditions, the appearance of the pertinent quantization nonlinearity in (5) may be modeled as

$$\mathbf{x}(k+1) = \mathcal{Q}[A \cdot \mathbf{x}(k)]. \tag{6}$$

Depending on whether the result of a product can be stored with full precision or whether quantization is performed immediately after each product is computed determines the effect of this nonlinearity. Considering (5) and noting that $\mathbf{x}(k) = \{x_i\} \in \Re^m$ and $A = \{a_{ij}\} \in \Re^{m \times m}$, we get the following:

If the products can be stored with full precision, that is, if a double-length accumulator is available,

$$\mathbf{x}(k+1) = \begin{pmatrix} \mathcal{Q}[\sum_{j=1}^{m} a_{1j} \cdot x_j(k)] \\ \vdots \\ \mathcal{Q}[\sum_{j=1}^{m} a_{mj} \cdot x_j(k)] \end{pmatrix} \tag{7}$$

and, on the other hand, if the product is quantized immediately after each product is performed, that is, if only a single-length accumulator is available,

$$\mathbf{x}(k+1) = \begin{pmatrix} \mathcal{Q}[a_{11} \cdot x_1(k)] + \mathcal{Q}[a_{12} \cdot x_2(k)] + \ldots + \mathcal{Q}[a_{1m} \cdot x_m(k)] \\ \vdots \\ \mathcal{Q}[a_{m1} \cdot x_1(k)] + \mathcal{Q}[a_{m2} \cdot x_2(k)] + \ldots + \mathcal{Q}[a_{mm} \cdot x_m(k)] \end{pmatrix} \tag{8}$$

Since q has been normalized to unity, noting (1), (7) and (8) may be expressed in a unified manner as

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k) + \mathbf{e}(k), \quad \text{with} \quad |e_i(k)| \leq N \cdot \varrho, \tag{9}$$

where $\mathbf{e}(k) = \{e_i(k)\} \in \Re^m$ and $e_i(k) \in \Re$. Note that, if (7) is applicable, $N = 1$; if (8) is applicable, $N = m$.

We note that, (9) is a description of a *linear* system driven by the bounded quantization error input $\mathbf{e}(k)$. Hence, we have in fact converted the nonlinear

systems in (7) and (8) into the linear system in (9). Now, the transfer function between $\mathbf{e}(k)$ and $\mathbf{x}(k)$ is

$$\frac{\mathbf{X}(z)}{\mathbf{E}(z)} = (z \cdot I - A)^{-1} \in \Re(z)_{m \times m}, \tag{10}$$

where $\mathbf{X}$ and $\mathbf{E}$ are the z-transforms of $\mathbf{x}$ and $\mathbf{e}$, respectively. This, when expanded, may be expressed as

$$\frac{\mathbf{X}(z)}{\mathbf{E}(z)} = \begin{pmatrix} H_{11}(z) & H_{12}(z) & \dots & H_{1m}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{m1}(z) & H_{m2}(z) & \dots & H_{mm}(z) \end{pmatrix}$$

where $H_{ij}(z) \in \Re(z)$. Hence,

$$X_i(z) = \sum_{j=1}^{m} H_{ij}(z) \cdot E_j(z), \; i = 1, 2, \dots, m, \tag{11}$$

where $\mathbf{X}(z) = \{X_i\}$ and $\mathbf{E}(z) = \{E_j\}$. Taking inverse z-transform of the above, we get

$$x_i(k) = \sum_{j=1}^{m} h_{ij}(k) * e_j(k), \; i = 1, 2, \dots, m,$$

where $h_{ij}(k)$ is the impulse response of $H_{ij}(z)$. Hence

$$x_i(k) = \sum_{j=1}^{m} \sum_{r=0}^{\infty} h_{ij}(r) \cdot e_j(k-r), \; i = 1, 2, \dots, m. \tag{12}$$

Combining (12) with the fact that $|e_j(k)| \leq N \cdot \varrho$, for $j = 1, 2, \dots, m$, we obtain

$$|x_i(k)| \leq N \cdot \varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|. \tag{13}$$

Eqn. (13) may now be used to provide upper bounds for each state vector x_i as follows:

$$M_i = N \cdot \varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|, \; i = 1, 2, \dots, m. \tag{14}$$

We realize that, in order to estimate a useful upper bound for x_i, we need to compute $\sum_{j=1}^{m}\sum_{k=0}^{\infty}|h_{ij}(k)|$ for a given filter. We address this now. Consider the transfer function $H_{ij}(z)$.

All poles of $H_{ij}(z)$ are distinct:

In this case, $H_{ij}(z)$ may be expressed as

$$H_{ij}(z) = K_{ij} + \frac{r_{ij}^{(1)}}{1 - P_1^{(1)}z^{-1}} + \ldots + \frac{r_{ij}^{(m)}}{1 - P_m^{(m)}z^{-1}},$$

where $r_{ij}^{(p)}, P_\ell^{(q)} \in \mathcal{C}$ and $K_{ij} \in \Re$, for $i, j, \ell, p, q = 1, 2, \ldots, m$. Taking the inverse z-transform, we have

$$h_{ij}(k) = K_{ij} \cdot \delta(k) + r_{ij}^{(1)}[P_1^{(1)}]^k + \ldots + r_{ij}^{(m)}[P_m^{(m)}]^k,$$

where $\delta(k)$ is the Dirac delta function. Therefore

$$\begin{aligned}\sum_{k=0}^{\infty}|h_{ij}(k)| &\leq \sum_{k=0}^{\infty}\{|K_{ij}||\delta(k)| + |r_{ij}^{(1)}|[|P_1^{(1)}|]^k + \ldots + |r_{ij}^{(m)}|[|P_m^{(m)}|]^k\} \\ &= |K_{ij}| + |r_{ij}^{(1)}|(1 - |P_1^{(1)}|)^{-1} + \ldots + |r_{ij}^{(m)}|(1 - |P_m^{(m)}|)^{-1}.\end{aligned}$$

This, when expanded, gives

$$\begin{aligned}\sum_{j=1}^{m}\sum_{k=0}^{\infty}|h_{ij}(k)| \leq \; & \textstyle\sum_{j=1}^{m}|K_{ij}| \quad + (1 - |P_1^{(1)}|)^{-1} \cdot \sum_{j=1}^{m}|r_{ij}^{(1)}| + \ldots \\ & + \ldots \qquad + (1 - |P_m^{(m)}|)^{-1} \cdot \sum_{j=1}^{m}|r_{ij}^{(m)}|,\end{aligned}$$

for $i = 1, 2, \ldots, m$. Hence

$$\begin{aligned}|x_i(k)| \leq \; & N \cdot \varrho \cdot \{\textstyle\sum_{j=1}^{m}|K_{ij}| + (1 - |P_1^{(1)}|)^{-1} \cdot \sum_{j=1}^{m}|r_{ij}^{(1)}| + \ldots \\ & \ldots + (1 - |P_m^{(m)}|)^{-1} \cdot \textstyle\sum_{j=1}^{m}|r_{ij}^{(m)}|\},\end{aligned} \tag{15}$$

for $i = 1, 2, \ldots, m$. Note that, convergence of the above is guaranteed due to linear stability of the digital filter.

Remark. The method adopted in [10] tends to be easier to implement and more general with regards to its capability of handling the presence of poles

of higher multiplicity. However, our experience has been that the technique described above often leads to lower upper bounds. Note that, the technique in [10] utilizes an interpretation that involves a cascade of first-order sections to obtain a bound for $|x_i|$; the technique above utilizes an interpretation that involves a parallel combination. Of course, no *one* technique will provide a lower bound for *all* situations. If computer cost is of concern, one can run both techniques and utilize the lower value of the bound.

$H_{ij}(z)$ contains a pole with multiplicity γ:

Let this pole of multiplicity γ be P. Then, $H_{ij}(z)$ may be expressed as

$$H_{ij}(z) = K_{ij} + \frac{r_{ij}^{(1)}}{(1 - Pz^{-1})} + \frac{r_{ij}^{(2)}}{(1 - Pz^{-1})^2} + \ldots + \frac{r_{ij}^{(\gamma)}}{(1 - Pz^{-1})^{\gamma}},$$

This analysis differs from the one given above for the general term

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})^{\zeta}}$$

where $\zeta = 2, 3, \ldots, \gamma$.

At this point, due mainly to its ease of implementation, we utilize the technique in [10] where the above expression is interpreted as a cascade of ζ first-order sections. For each first-order section, the inverse z-transform is taken using the theory outlined in the distinct pole case. Consider

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})^{\zeta}} = \frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})(1 - Pz^{-1}) \ldots (1 - Pz^{-1})}.$$

Taking the inverse z-transform, we get

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})(1 - Pz^{-1}) \ldots (1 - Pz^{-1})} = r_{ij}^{(\zeta)} \cdot \left[\sum_{k=0}^{\infty} |P|^k\right]^{\zeta} = r_{ij}^{(\zeta)} \cdot \left[\frac{1}{1 - |P|}\right]^{\zeta}.$$

This expression is now substituted for the pole of multiplicity γ.

Lemma 1: The zero input response of the state $\mathbf{x}(k)$ of the digital filter described by eqn (7) or (8) is periodic. Its period T satisfies

$$T \leq \prod_{i=1}^{m} (2 \cdot \hat{M}_i + 1) = T_{max}, \tag{16}$$

where $\hat{M}_i$ is the largest integer not more than M_i in eqn (14).

Proof: Consider eqn (7) or (8). The steady-state solution of each state $x_i(k)$ will satisfy

$$|x_i(k)| \leq M_i, \quad \forall k,\ i = 1, 2 \ldots, m.$$

Under fixed-point arithmetic, $\mathbf{x}(k) \in \mathcal{Z}^m$, and hence,

$$|x_i(k)| \leq \hat{M}_i,\ \forall k,\ i = 1, 2 \ldots, m.$$

$x_i(k)$ can therefore take only a finite number of values, namely, $(2 \cdot \hat{M}_i + 1)$. As a result of this, $\mathbf{x}(k)$ can take only a finite number of values, namely,

$$\prod_{i=1}^{m} (2 \cdot \hat{M}_i + 1).$$

Note that, the current state vector $\mathbf{x}(k)$ uniquely determines the next state vector $\mathbf{x}(k+1)$ through the function $\mathcal{Q}[\cdot]$. Thus, $\mathbf{x}(k)$ must be periodic in k. Its period is in fact bounded by

$$T_{max} = \prod_{i=1}^{m} (2 \cdot \hat{M}_i + 1). \tag{17}$$

□

We now have bounds on the amplitude as well as the period on the possible limit cycles. This information will be invaluable for developing our search algorithm.

IV Algorithm Description and Its Computational Aspects

In this section, we formulate the theoretical basis for the algorithm and discuss some of its computational aspects.

Definition 1: The digital filter realization in (9) is said to be globally asymptotically stable (g.a.s.) if and only if, for any initial state $\mathbf{x}(0) \in \mathcal{Z}^m$ with

$\|\mathbf{x}(0)\|_\infty \leq B$, where $B \in \mathcal{Z}_+$, there exists $L \in \mathcal{Z}_+$ such that $\mathbf{x}(k) = \mathbf{0}$ for $k \geq L$.

Remark. Typically, g.a.s. is taken to hold when $\mathbf{x}(k) \rightarrow \mathbf{0}$ as $k \rightarrow \infty$ (under the conditions above). However, due to the finite wordlength available in each register, the digital filter behaves as a finite state machine, and Definition 1 suffices.

Lemma 2: Consider $\eta > 0$ and any initial state vector $\mathbf{x}(0)$ such that

$$|x_i(0)| \leq B_i, \quad \text{for} \quad i = 1, 2, \ldots, m,$$

with $B_i > \hat{M}_i$, for $i = 1, 2, \ldots, m$. Then, there exists a sufficiently large positive number $\mathcal{L}$ such that the digital filter in (7) or (8) satisfies

$$|x_i(k)| \leq \hat{M}_i + \eta, \ \forall k \geq \mathcal{L},$$

for $i = 1, 2, \ldots, m$.

Proof: Since the eigenvalues of A are assumed to lie inside the unit circle in the complex plane, the digital filter in eqn. (9) is in fact g.a.s. Hence, eqn. (9) will yield a set of nonhomogeneous linear shift-invariant difference equations which will have its solution in two parts: A steady-state solution $\mathbf{s}(k)$ and a transient solution $\mathbf{t}(k)$. Clearly, with g.a.s., given $\eta > 0$, we can choose k sufficiently large, say, $k \geq \mathcal{L}$, such that

$$\max |t_i(k)| < \eta, \quad \text{for} \quad i = 1, 2, \ldots, m.$$

Since $\hat{M}_i \in \mathcal{Z}_+$, for $k \geq \mathcal{L}$, $\hat{M}_i + \eta$ will therefore act as a true upperbound for $x_i(k)$ in eqn. (9). □

Hence, it suffices to check the state vectors in the set $\mathcal{S}^{(0)}$, where

$$\mathcal{S}^{(0)} = \left\{ \mathbf{x}(k) \in \mathcal{Z}^m \,\middle|\, |x_i(k)| \leq \hat{M}_i, \ i = 1, 2, \ldots, m \right\}, \tag{18}$$

to see if they are mapped to the zero vector by eqn. (9) after a finite number of mappings.

Computational Aspects

The computations within the algorithm are carried out in two stages. Initially, all vectors $\mathbf{x}(k) \in \mathcal{S}^{(0)}$ which map to $\mathbf{0}$ in less than T_{max} recursions—(after all, if limit cycles exist, the maximum period is T_{max})—are eliminated from $\mathcal{S}^{(0)}$ as they are now known to be stable. The remaining vectors in $\mathcal{S}^{(0)}$ are then further checked for convergence (see Section B).

Section A. Consider the set $\mathcal{V}^{(1)}$, where

$$\mathcal{V}^{(1)} = \left\{\mathbf{x}(k) \in \mathcal{S}^{(0)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] = \mathbf{0}\right\}, \tag{19}$$

Hence, $\mathcal{V}^{(1)}$ consists of all the vectors $\mathbf{x}(k) \in \mathcal{S}^{(0)}$ that map to $\mathbf{0}$ in one and only one iteration of equation (7) or (8). Note that, any other stable vector in $\mathcal{S}^{(0)}$ must map to $\mathcal{V}^{(1)}$ prior to reaching $\mathbf{0}$. Hence, for further computations, we form

$$\mathcal{S}^{(1)} = \mathcal{S}^{(0)} \setminus \mathcal{V}^{(1)}. \tag{20}$$

Note that, $\mathcal{K}[\mathcal{S}^{(1)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \mathcal{K}[\mathcal{V}^{(1)}]$. In fact, one immediately notices that $\mathcal{K}[\mathcal{S}^{(0)}] = T_{max}$.

Furthermore, any vector in $\mathcal{S}^{(1)}$ which is mapped to $\mathcal{V}^{(1)}$ by (7) or (8) in one iteration will also converge to $\mathbf{0}$. Hence, we form the set $\mathcal{V}^{(2)}$, where

$$\mathcal{V}^{(2)} = \left\{\mathbf{x}(k) \in \mathcal{S}^{(1)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] \in \mathcal{V}^{(1)}\right\}. \tag{21}$$

Hence, $\mathcal{V}^{(2)}$ consists of all the vectors $\mathbf{x}(k) \in \mathcal{S}^{(1)}$ that map to $\mathbf{0}$ in exactly two iterations of equation (7) or (8). Hence, for further computations, we form

$$\mathcal{S}^{(2)} = \mathcal{S}^{(1)} \setminus \mathcal{V}^{(2)}. \tag{22}$$

Note that, $\mathcal{K}[\mathcal{S}^{(2)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \mathcal{K}[\mathcal{V}^{(1)}] - \mathcal{K}[\mathcal{V}^{(2)}]$.

Likewise, we get the following sets: For $L = 1, 2, \ldots, T_{max}$,

$$\mathcal{V}^{(L)} = \left\{\mathbf{x}(k) \in \mathcal{S}^{(L-1)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] \in \mathcal{V}^{(L-1)}\right\}, \tag{23}$$

and

$$\mathcal{S}^{(L)} = \mathcal{S}^{(L-1)} \setminus \mathcal{V}^{(L)}. \tag{24}$$

Note that, $\mathcal{K}[\mathcal{S}^{(L)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \sum_{i=1}^{L} \mathcal{K}[\mathcal{V}^{(i)}]$.

The conditions under which this construction is terminated and their implications are as follows:
(1) If

$$\mathcal{K}[\mathcal{S}^{(L)}] = \emptyset, \quad \text{for some} \quad L = 1, 2, \ldots, T_{max} - 1, \tag{25}$$

all vectors in $\mathcal{S}^{(0)}$ are convergent.

(2) If

$$\mathcal{K}[\mathcal{V}^{(L)}] = \emptyset, \quad \text{for some} \quad L = 1, 2, \ldots T_{max}, \tag{26}$$

then

$$\mathcal{S}^{(i)} = \mathcal{S}^{(L-1)}, \quad \text{for} \quad i = L, L+1, \ldots, T_{\max}. \tag{27}$$

Under this situation, the remaining vectors in $\mathcal{S}^{(L-1)}$–there are $\mathcal{K}[\mathcal{S}^{(L-1)}]$ of them–will be further checked for convergence (see Section B).

Remark. Upon a little reflection, one notices that $\mathcal{V}^{(T_{max})}$ must either be empty or contain one and only one vector from $\mathcal{S}^{(0)}$.

Section B. Although the reverse mapping procedure outline above reduces the computational complexity considerably, it may not capture all the vectors in $\mathcal{V}^{(L)}$, $L = 1, 2, \ldots, T_{max}$, that map to **0** within T_{max} iterations. This is due to the fact that, there may be vectors in $\mathcal{V}^{(L)}$ that map to **0** through a vector not belonging to $\mathcal{S}^{(0)}$! Hence, when encountered with condition (2) above, convergence of each remaining vector in $\mathcal{S}^{(L-1)}$ is determined by checking whether it is mapped to **0** in less than T_{max} through either (7) or (8), whichever is applicable. This exhaustive technique is in fact an extension of that given in [10] to digital filters represented in their state-space realization. However, we must emphasize the significant computational advantage gained by first invoking the reverse mapping construction procedure in Section A.

Assuming condition (2) has occurred, let

$$\mathcal{S}^{(L)} = \left\{ \mathbf{x}_i^{(L)};\ i = 1, 2, \ldots, \mathcal{K}[\mathcal{S}^{(L)}] \right\}. \tag{28}$$

Note that, when condition (2) has occurred, from (27), $\mathcal{S}^{(L-1)} = \mathcal{S}^{(L)}$. For each vector $\mathbf{x}_i^{(L)} \in \mathcal{S}^{(L)}$, construct the orbit $\mathcal{O}_i^{(L)}$ consisting of all state vectors

$\mathbf{x}_i^{(L)}(j)$, for $j = 1, 2, \ldots, T_{max}$, that are consecutively generated by (7) or (8) (whichever is applicable) with $\mathbf{x}_i^{(L)}$ as the initial state, that is, $\mathbf{x}_i^{(L)} = \mathbf{x}_i^{(L)}(0)$.

For each $i = 1, 2, \ldots, \mathcal{K}[\mathcal{S}^{(L)}]$, the conditions under which the construction of each orbit $\mathcal{O}_i^{(L)}$ is terminated and their implications are as follows:
(1) If

$$\mathbf{x}_i^{(L)}(j) = 0, \quad \text{for some} \quad j = 1, 2, \ldots, T_{max}, \tag{29}$$

then $\mathbf{x}_i^{(L)}$ together with each vector in the orbit $\mathcal{O}_i^{(L)}$ is convergent.
(2) If

$$\mathbf{x}_i^{(L)}(j) = \mathbf{x}_i^{(L)}(k), \quad \text{for} \quad j \neq k, \tag{30}$$

then $\mathbf{x}_i^{(L)}$ gives rise to limit cycles.

Remark. These are in fact the only conditions that can occur when either (7) or (8) generate the orbit.

V Perturbation of Filter Coefficient Matrix

In constructing the region of g.a.s. in the coefficient space, perturbations incurred in storing each filter coefficient must also be considered. Such perturbations are typically due to finite wordlength effects that require rounding or truncation of the true coefficient value.

The algorithm described in the previous section provides information regarding g.a.s. of a given filter with a nominal coefficient matrix $A = \{a_{ij}\} \in \Re^{m \times m}$. Once this is done, we now consider a small perturbation Δa_{ij} of each coefficient about its nominal value a_{ij}. However, for a given state vector $\mathbf{x}(k)$, this perturbation may not necessarily alter the next state $\mathbf{x}(k+1)$ obtained since it is entirely possible that

$$\mathbf{x}(k+1) = \mathcal{Q}[(A + \Delta A) \cdot \mathbf{x}(k)] = \mathcal{Q}[A \cdot \mathbf{x}(k)], \tag{31}$$

where $\Delta A = \{\Delta a_{ij}\} \in \Re^{m \times m}$.

Depending on the number of quantizers per row, that is, depending on whether a double- or single-length accumulator is available, (31) is interpreted differently.

Double-length accumulator

It is evident that the upper bound $\hat{M}_i$ estimated for the nominal value of the coefficient matrix $\{a_{ij}\} \in \Re^{m\times m}$ will no longer be valid for a perturbed system $\{a_{ij}+\Delta a_{ij}\} \in \Re^{m\times m}$. If for a filter with a nominal coefficient matrix $\{a_{ij}\}$, the upper bound valid for all systems described by the coefficient matrices $\{a_{ij} + \Delta a_{ij}\}$, be $\tilde{M}_\nu$ $(\nu = 1, 2, \ldots, m)$, this then could be used to estimate the robustness region as explained below. The choice of $\tilde{M}_\nu$ is critical. To determine the robustness region we need an estimate for $\tilde{M}_\nu$ which will be valid for all systems with $\{\Delta a_{ij}\}$ perturbations, an vise versa. From eqn. (32) we see that $\mathcal{G}$ can only be determined after a suitable $\tilde{M}_\nu$ is chosen. If the chosen $\tilde{M}_\nu$ is large the region $\mathcal{G}$ will decrease. If a small value for $\tilde{M}_\nu$ is chosen the region $\mathcal{G}$ will become large and $\tilde{M}_\nu$ will no longer be a valid upper bound for systems $\{a_{ij} + \Delta a_{ij}\}$. In such a situation the robustness region will be an intersection of the computed $\mathcal{G}$ and the region formed by $\{a_{ij}+\Delta a_{ij}\}$ in the parameter space where $\tilde{M}_\nu$ is still a valid upper bound.

Consider the *robustness region*

$$\begin{aligned}\mathcal{G} \;=\; \{\Delta a_{\nu j} | \mathcal{Q}\left[\sum_{j=1}^{m}(a_{ij}+\Delta a_{ij})\cdot x_j(k)\right] = \mathcal{Q}\left[\sum_{j=1}^{m} a_{ij}\cdot x_j(k)\right],\\ \text{for } i = 1,2,\ldots,m, \text{ and } \forall \mathrm{x}(k) \in \tilde{\mathcal{S}}\},\end{aligned} \tag{32}$$

where

$$\tilde{\mathcal{S}} = \left\{\mathrm{x}\,\middle|\, |x_\nu| \le \tilde{M}_\nu,\ \ \nu = 1,2,\ldots,m;\ \mathrm{x} \in \mathcal{Z}^m\right\}.$$

Here, $\tilde{M}_\nu$ is an upper bound valid for *all* filters described by the coefficient matrices $\{a_{ij} + \Delta a_{ij}\}$.

Hence we can assume that $\tilde{\mathcal{S}}$ is valid for all systems described by (7) with coefficients $\{a_{ij} + \Delta a_{ij}\}$.

To proceed, it is convenient to identify the discontinuities associated with the nonlinearity $\mathcal{Q}[\cdot]$.
For sign-magnitude roundoff,

$$\mathcal{D}_r = \left\{b_r \in \Re \mid b_r = r + \frac{1}{2},\ r \in \mathcal{Z}\right\}; \tag{33}$$

for sign-magnitude truncation quantization,

$$\mathcal{D}_{mt} = \{b_r \in \Re \mid b_r = r,\ r \in \mathcal{Z} \setminus \{0\}\}\,; \tag{34}$$

for two's complement truncation quantization,

$$\mathcal{D}_{two} = \{b_r \in \Re \mid b_r = r,\ r \in \mathcal{Z}\}\,. \tag{35}$$

For each $\mathbf{x} \in \tilde{\mathcal{S}}$, a region $\mathcal{G}_{\mathbf{x}}$ corresponding to the robustness region in eqn. (32) applicable to the pertinent quantization schemes in (33), (34), or (35) is defined. Let the region corresponding to the ν-th state x_ν of $\mathbf{x}$ be $\mathcal{G}_{\mathbf{x}}^{(\nu)}$. Then, we have the following:

For sign-magnitude roundoff quantization,

$$\mathcal{G}_{\mathbf{x}}^{(\nu)} = \left\{ \begin{array}{ll} \{\Delta a_{\nu j} | & b_{r-1} - \sum_{j=1}^m a_{\nu j}x_j \le \sum_{j=1}^m \Delta a_{\nu j}x_j < b_r - \sum_{j=1}^m a_{\nu j}x_j\} \\ & \text{for } b_{r-1} \le \sum_{j=1}^m a_{\nu j}x_j < b_k \text{ and } r \ge 1 \\ \{\Delta a_{\nu j} | & b_{r-1} - \sum_{j=1}^m a_{\nu j}x_j < \sum_{j=1}^m \Delta a_{\nu j}x_j \le b_r - \sum_{j=1}^m a_{\nu j}x_j\} \\ & \text{for } b_{r-1} < \sum_{j=1}^m a_{\nu j}x_j \le b_k \text{ and } r \le -1 \\ \{\Delta a_{\nu j} | & b_{-1} - \sum_{j=1}^m a_{\nu j}x_j < \sum_{j=1}^m \Delta a_{\nu j}x_j < b_0 - \sum_{j=1}^m a_{\nu j}x_j\} \\ & \text{for } b_{-1} < \sum_{j=1}^m a_{\nu j}x_j < b_0 \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_r; \tag{36}$$

for sign-magnitude truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(\nu)} = \left\{ \begin{array}{ll} \{\Delta a_{\nu j} | & b_{r-1} - \sum_{j=1}^m a_{\nu j}x_j \le \sum_{j=1}^m \Delta a_{\nu j}x_j < b_r - \sum_{j=1}^m a_{\nu j}x_j\} \\ & \text{for } b_{r-1} \le \sum_{j=1}^m a_{\nu j}x_j < b_k \text{ and } r \ge 2 \\ \{\Delta a_{\nu j} | & b_{r-1} - \sum_{j=1}^m a_{\nu j}x_j < \sum_{j=1}^m \Delta a_{\nu j}x_j \le b_r - \sum_{j=1}^m a_{\nu j}x_j\} \\ & \text{for } b_{r-1} < \sum_{j=1}^m a_{\nu j}x_j \le b_k \text{ and } r \le -1 \\ \{\Delta a_{\nu j} | & b_{-1} - \sum_{j=1}^m a_{\nu j}x_j < \sum_{j=1}^m \Delta a_{\nu j}x_j < b_{+1} - \sum_{j=1}^m a_{\nu j}x_j\} \\ & \text{for } b_{-1} < \sum_{j=1}^m a_{\nu j}x_j < b_{+1} \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{mt}; \tag{37}$$

for two's complement truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(\nu)} = \left\{ \begin{array}{ll} \{\Delta a_{\nu j} | & b_{r-1} - \sum_{j=1}^{m} a_{\nu j} x_j \leq \sum_{j=1}^{m} \Delta a_{\nu j} x_j < b_r - \sum_{j=1}^{m} a_{\nu j} x_j \} \\ & \text{for } b_{r-1} \leq \sum_{j=1}^{m} a_{\nu j} x_j < b_k \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{two}. \tag{38}$$

For a particular state in the vector $\mathbf{x}$, the region can be computed using eqns.(36) , (37) and (38). For all $\mathbf{x} \in \tilde{\mathcal{S}}$ the total robustness region is given by

$$\mathcal{G} = \bigcap_{\mathbf{x} \in \tilde{\mathcal{S}}} \mathcal{G}_{\mathbf{x}}. \tag{39}$$

From (36) , (37) and (38), we may estimate suitable values for the region of robustness for each quantization scheme. For the twos complement truncation quantization scheme from eqn. (38), let

$$\sum_{j=1}^{m} \Delta a_{\nu j} x_j \leq \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{\nu j} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{\nu j} x_j| \right\} \tag{40}$$

The left hand side of (40) will be given by,

$$\left| \sum_{j=1}^{m} \Delta a_{\nu j} x_j \right| \leq \sum_{j=1}^{m} |\Delta a_{\nu j}| \cdot |x_j| \tag{41}$$

If M_{max} is the maximum value of $\tilde{M}_\nu$ for all ν, then $|x_j| < M_{max}$ for all j is true. eqn.(41) then can be simplified to,

$$\left| \sum_{j=1}^{m} \Delta a_{\nu j} x_j \right| \leq M_{max} \cdot \sum_{j=1}^{m} |\Delta a_{\nu j}| \tag{42}$$

If we estimate the perturbations $\Delta a_{\nu j}$ such that

$$M_{max} \cdot \sum_{j=1}^{m} |\Delta a_{\nu j}| < \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{\nu j} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{\nu j} x_j| \right\} \tag{43}$$

$$\sum_{j=1}^{m} |\Delta a_{\nu j}| < \frac{1}{M_{max}} \cdot \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{\nu j} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{\nu j} x_j| \right\} \tag{44}$$

Using the approximation in eqn. (44) we can estimate a region $\hat{\mathcal{G}}$, where $\hat{\mathcal{G}}$ is given by,

$$\hat{\mathcal{G}} = \{\Delta a_{\nu j} | \ \|\Delta a_{\nu j}\|_1 < \frac{1}{M_{max}} \cdot \min_{x \in \tilde{\mathcal{S}}} \{|b_r - \sum_{j=1}^{m} a_{\nu j} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{\nu j} x_j|\}\}. \tag{45}$$

Clearly, $\hat{\mathcal{G}} \subset \mathcal{G}$.
From eqn.(45) it is observed that in a degenerate case $\hat{\mathcal{G}}$ may only contain the zero perturbation vector. Generally not all perturbations $\Delta a_{\nu j}$ are zero. To find the region of robustness for the sign-magnitude truncation and roundoff schemes a similar analysis can be carried out.

In the case of sign-magnitude roundoff,

$$\min_{x \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{\nu j} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{\nu j} x_j| \right\} < \frac{1}{2} \tag{46}$$

Then the maximum perturbation region, $\hat{\mathcal{G}}_{max}$ is given by,

$$\hat{\mathcal{G}}_{max} = \left\{ \Delta a_{\nu j} | \ \|\Delta a_{\nu j}\|_1 < \frac{1}{2\hat{M}_\nu}; \quad \nu = 1, 2 \ldots, m \right\} \tag{47}$$

Where $\hat{M}_\nu$ for $\nu = 1, 2 \ldots, m$ are the upper bounds computed for the nominal value $\{a_{ij}\}$. Since $M_{max} \geq \tilde{M}_\nu \geq \hat{M}_\nu$, $\hat{\mathcal{G}} \subset \hat{\mathcal{G}}_{max}$. Therefore eqn. (47) can be used to obtain a valid upper bound in this case.

To compute the region described above requires a complex algorithm, the computation load can be greatly reduced by following the guidelines given below.

Initially if vectors with comparatively higher upper bounds are used in the computation, the robustness region will initially converge faster and the number of added vertices to the formation of the total region will be very little due to vectors with comparatively lower upper bounds. It is observed from eqn.(32) that when the upper bound is small the region described by $\mathcal{G}$ tends to become larger.

Observation:

If $\tilde{M}_\nu < 1$ for all ν we observe that $\tilde{\mathcal{S}}$ will only contain $\mathbf{0}$. It follows that for a digital filter implementation given by eqn. (7)

$$\varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| < \tilde{M}_\nu < 1 \tag{48}$$

for $\nu = 1, 2 \ldots, m$
Since,

$$\sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| > 1 \tag{49}$$

Eqn. (48) can only be satisfied if $\varrho < 1$. For sign magnitude roundoff quantization $\varrho = 1/2$, therefore from eqn. (48),

$$\sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| < 2 \tag{50}$$

Therefore we conclude that a digital filter in double length accumulator environment satisfying eqn. (50), is globally asymptotically stable.

Single-length accumulator

If there are m quantizers per row as in (8), robustness region is defined in the following manner:

$$\mathcal{G} = \left\{ \Delta a_{ij} | \mathcal{Q}\left[\sum_{j=1}^{m} (a_{ij} + \Delta a_{ij}) \cdot x_j\right] = \mathcal{Q}\left[\sum_{j=1}^{m} (a_{ij} \cdot x_j)\right], \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1 \right\}. \tag{51}$$

As in the double length accumulator implementation we define a upper bound valid for all systems given by eqn. (8), and define a set $\tilde{\mathcal{S}}_1$.

$$\tilde{\mathcal{S}}_1 = \left\{ \mathbf{x} \middle| x_i \leq \tilde{M}_i \,; \quad i = 1, 2, \ldots, m; \quad \mathbf{x} \in \mathcal{Z}^m \right\} \tag{52}$$

Where $\tilde{M}_i$ is an upper bound valid for all filters described by the coefficient matrix $\{a_{ij} + \Delta a_{ij}\}$. Let the robustness region corresponding to element a_{ij} in the coefficient matrix for a particular state vector $\mathbf{x}$ be $\mathcal{G}_{\mathbf{x}}^{(i,j)}$. Then, we have the following:

For sign-magnitude roundoff quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i,j)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j \leq \Delta a_{ij} \cdot x_j < b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} \leq a_{ij} \cdot x_j < b_k \text{ and } r \geq 1 \\ \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j \leq b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} < a_{ij} \cdot x_j \leq b_k \text{ and } r \leq -1 \\ \{\Delta a_{ij} | & b_{-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j < b_0 - a_{ij} \cdot x_j\} \\ & \text{for } b_{-1} < a_{ij} \cdot x_j < b_0 \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{r,} \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1; \tag{53}$$

for sign-magnitude truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i,j)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j \leq \Delta a_{ij} \cdot x_j < b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} \leq a_{ij} \cdot x_j < b_k \text{ and } r \geq 2 \\ \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j \leq b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} < a_{ij} \cdot x_j \leq b_k \text{ and } r \leq -1 \\ \{\Delta a_{ij} | & b_{-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j < b_{+1} - a_{ij} \cdot x_j\} \\ & \text{for } b_{-1} < a_{ij} \cdot x_j < b_{+1} \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{mt} \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1 \tag{54}$$

For two's complement truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i,j)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j \leq \Delta a_{ij} \cdot x_j < b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} \leq a_{ij} \cdot x_j < b_k \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{two} \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1. \tag{55}$$

Hence g.a.s. can be gaurenteed for the region

$$\mathcal{G} = \bigcap_{\forall(i,j)} \mathcal{G}_{\mathbf{x}}^{(i,j)}. \tag{56}$$

Using a similar argument as in the case of a double-length accumulator, we can estimate a region of robustness for each quantization scheme. The estimated region for the sign magnitude roundoff will be given by,

$$\hat{\mathcal{G}} = \left\{ \Delta a_{ij} \left| |\Delta a_{ij}| < \frac{1}{2\tilde{M}_i} \; ; \quad i = 1, 2, \ldots, m \right. \right\} \tag{57}$$

VI Some Examples

In this section the proposed search algorithm is applied to a dense grid in the coefficient space to obtain the total global asymptotic stability region for a digital filter with zero input. The dense grid will provide a reasonably good approximation to the g.a.s region, since it is not possible to consider all points in the linear stability region. Note that each point in the coefficient space is associated with a neighborhood where the filter is stable. A 10 Bit wordlength is assume for all computations, therefore the filter coefficients are quantized to a multiple of 2^{-10}. Within the linear stability region dark areas indicate points where limit cycles of some period exists. It should be noted that the linear stability region does not have a common boundary with the global asymptotic stability region obtained through this algorithm. Therefore in all figures, the boundary line which delimits the stability region from the unstable region does not belong to the stability region.

The most commonly encountered quantization schemes are analyzed, they are namely, sign magnitude roundoff quantization scheme, sign-magnitude truncation quantization scheme and the two's complement truncation quantization scheme. In all quantization schemes the single- and the double-length accumulator implementation results are provided. All results are provided for the $\{a_{ij}\} \in \Re^{2\times 2}$ coefficient matrix. All existing results for the named quantization schemes were verified. For a direct form digital filter in state space formulation (the coefficient matrix is given by eqn.(58)

$$A = \begin{bmatrix} 0 & 1 \\ a_2 & a_1 \end{bmatrix} \tag{58}$$

Figure.(1a) shows the region obtained by the proposed algorithm the sign magnitude roundoff quantization scheme in an double length accumulator environment.

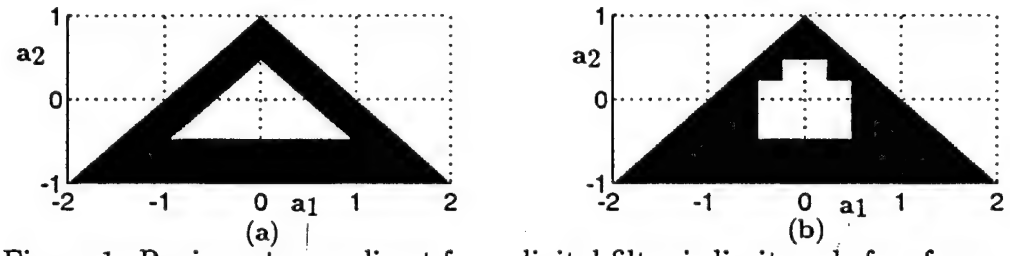


Figure 1: Region where a direct form digital filter is limit cycle free for cases (a) Double length accumulator. (b) single length accumulator

The region obtained is identical to the results given in [10]. For the same quantization scheme and single length accumulator the region obtained is given in Figure.(1b). The region matches exactly with the ones found in [10]. The regions for the two's complement and the sign magnitude truncation schemes were also verified. The regions obtained by the proposed algorithm matches with the regions given in [10].

Results for minimum norm realization of digital filters .

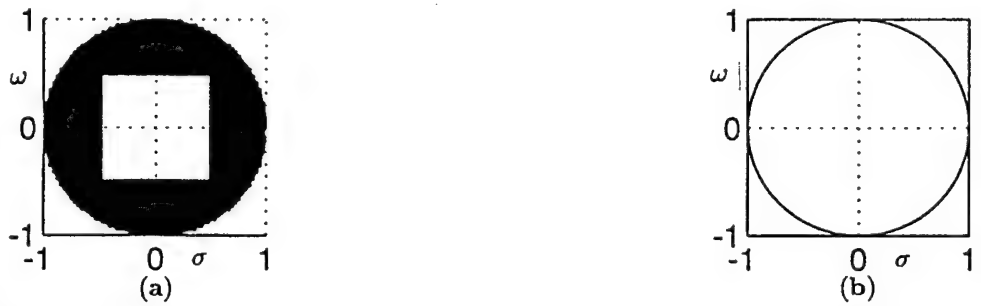


Figure 2: The region where a minimum norm realization of a digital filter is free of limit cycles for double- and single-length accumulator environments. (a) sign magnitude roundoff (b) sign magnitude Truncation

The stability of digital filters in its minimum norm form for the coefficient matrix, $\{a_{ij}\} \in \Re^{2\times 2}$ case was also investigated. The coefficient matrix is

given by eqn.(59).

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix} \tag{59}$$

The results for the sign magnitude roundoff scheme for the single- and the double-length accumulator environment is given in Figure.(2a) This region matches with the region given in [7]. The stable region for the sign magnitude truncation scheme in a single length or double length accumulator environment spans the entire region where $\sigma^2 + \omega^2 < 1$. results are given in Figure.(2b). This supports the results obtained in [ ].

For the two's complement truncation quantization, with double length accumulator the global asymptotic region is given in Figure.(3a). This supports and also improves on the previously known results given in [19]. To the authors knowledge no previous results are available for the two's complement truncation quantization in a single length accumulator environment. The region of global asymptotic stability is summarized in Figure.(3b). Note that

Figure 3: Region where Two's complement truncation implementation of a minimum norm digital filter is limit cycle free (a) Single length accumulator (b) double length accumulator

for the Two's complement quantization scheme in a double length accumulator environment, series of points extend from the stability region into the instability region such that,

$$\sigma < 0 \quad \text{and} \quad \omega = \pm\sigma \tag{60}$$

The following coefficient matrix can be cited as an example,

$$A = \begin{bmatrix} -\frac{672}{1024} & \frac{672}{1024} \\ -\frac{672}{1024} & -\frac{672}{1024} \end{bmatrix} \tag{61}$$

VII Conclusion

A new algorithm capable of determining global asymptotic stability of any fixed point digital filter represented in its state space formulation, under zero input conditions has been presented. The search algorithm is independent of the type of nonlinearity, the number of nonlinearities and it has been generalized to handle a digital filter of order m in its state space represented form.

The proposed algorithm is found to provide tighter bounds on the amplitude of limit cycles in most cases, and it will always determine the stability or instability of a particular digital filter. Significant improvement over the existing results for the two's complement truncation schemes in both single- and double length accumulator environments have been presented.

The current research is directed towards the following problems.
(1) Establishing regions within which limit cycles of a pre-specified period exists.
(2) Establish regions within which limit cycles that are under a pre-specified bound exist.
(3) Extension of the algorithm for δ-operator formulated systems. In Fixed-point arithmetic it is known that such systems always exhibit limit cycle behavior [20]. Therefore in actual applications the regions similar to the ones mentioned in items (1) and (2) may be of importance.

References

1. E.I. Jury and B.W. Lee, "The absolute stability of systems with many nonlinearities," *Automat. Remote Contr.,* vol 26, no. 6, pp. 943-961, 1965.

2. W. Barnes and A.T. Fam, "Minimum norm recursive digital filters that are free of overflow oscillations," *IEEE Trans. Circ. Syst.,* vol. CAS-24, no. 10, pp. 569-574, Oct. 1977.

3. W.L. Mills, C.T. Mullis, and R.A. Roberts, "Digital filter realizations without overflow oscillations," *Proc. 1978 IEEE Int. Conf. Acoust., Speech, Sig. Proc.*, pp. 71-74, 1978.

4. T. Claasen, W.F.G. Mecklenbräuker, and J.B.H. Peek, "Frequency domain criteria for the absence of zero-input limit cycles in nonlinear discrete-time systems with applications to digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-22, no. 3, pp. 232-239, Mar. 1974.

5. E.D. Garber, "Frequency criteria for the absence of periodic responses," *Automat. Remote. Contr.*, vol. 28, no. 11, pp. 1776-1780, 1967.

6. V. Singh, "An extension to Jury-Lee's criterion for the stability analysis of fixed-point digital filters designed with two's complement arithmetic," *IEEE Trans. Circ. Syst.*, vol. CAS-33, no. 3, p. 355, Mar. 1986.

7. K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions—Part I: Direct form and coupled form filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 2, pp. 113-131, Feb. 1985.

8. K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions—Part II: Wave digital filters and lattice digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 2, pp. 132-142, Feb. 1985.

9. A.N. Michel and R.K. Miller, "Stability analysis of discrete time interconnected systems via computer generated Lyapunov functions with applications to digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 8, pp. 737-753, Aug. 1985.

10. P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed-point digital filters," *IEEE Trans. Sig. Proc.*, vol. 39. no. 11, pp. 2400-2409, Nov. 1991.

11. J.F. Kaiser, "Some special practical considerations in the realization of linear digital filters," *Proc. 3rd Allerton Ann. Conf. Circ. Syst. Theory*, pp. 100-104, 1965.

12. T. Bose and D.P. Brown, "Limit cycles in zero input digital filters due to two's complement quantization," *IEEE Trans. Circ. Syst.,* vol. CAS-37, no. 4, pp. 568-571, Month April 1990.

13. A. Lepschy, G.A. Mian and U. Viaro, "Effects of quantization in second-order fixed-point digital filters with two's complement truncation quantization," *IEEE Trans. Circ. Syst.,* vol. CAS-35, no. 4, pp. 461-466, April 1988.

14. Trân-Thông and B. Liu, "Limit cycles in the combination implementation of digital filters," *IEEE Trans. Acoust., Speech, Sig. Proc.,* vol. ASSP-24, no. 3, pp. 248-256, Feb. 1976.

15. Trân-Thông and B. Liu, "A contribution to the stability analysis of second-order direct-form digital filters with magnitude truncation," *IEEE Trans. Acoust., Speech, Sig. Proc.,* vol. ASSP-35, no. 8, pp. 1207-1210, Aug. 1987.

16. Trân-Thông and B. Liu, "Parameter space quantization in fixed-point digital filters," *Electron. Lett.,* vol. 22, no. 7, pp. 384-386, Mar. 1986.

17. Trân-Thông and B. Liu, "Parameter plane quantization induced by signal quantization in second-order fixed-point digital filters with one quantizer," *Sig. Proc.,* vol. 14, no. ?, pp. 103-106, Jan. 1988.

18. Trân-Thông and B. Liu, "Zero-input limit cycles and stability in second order fixed point digital filters with two magnitude truncation quantizers," *Circ., Syst., Sig. Proc.,* vol. 8, no. 4, pp. ?, 1989.

19. T. Bose, "Stability of digital filters implemented with Two's complement truncation quantization," *IEEE Trans. Sig. Proc.,* vol. 40. no. 01, pp. 24-31, Jan. 1992.

20. K. Premaratne, P. H. Bauer, "Limit cycles and asymptotic stability of delta-operator formulator discrete time systems implemented in fixed point arithmetic," *Proc. IEEE intl. Symp. on Circ. Syst. ISCAS '94 London UK*, pp. 461-464 ,May-June 1994.

Limit Cycles in Delta-Operator Formulated 1-D and M-D Discrete-Time Systems with Fixed-Point Arithmetic

Peter H. Bauer
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556

Kamal Premaratne
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL. 33124

ABSTRACT

In this paper, the problem of global asymptotic stability of δ-operator formulated one-dimensional (1-D) and multi-dimensional (m-D) discrete-time systems is analyzed for the case of fixed point implementations. It is shown that the free response of such a system tends to produce incorrect equilibrium points if conventional quantization arithmetic schemes such as truncation or rounding are used. Explicit necessary conditions for global asymptotic stability are derived in terms of the sampling period. These conditions demonstrate that, in almost all cases, fixed-point arithmetic does not allow for global asymptotic stability in δ-operator formulated discrete-time systems that use a short sampling time. This is true for the 1-D as well as the m-D case.

I. INTRODUCTION

Discrete-time systems formulated in terms of the incremental difference operator (or, δ-operator) have recently been receiving considerable attention in the technical literature [1-4]. Most of this work focuses on the superior performance of the δ-operator under finite wordlength conditions when compared with the shift-operator (or, q-operator). In particular, investigations of coefficient sensitivity and quantization noise properties have revealed that δ-operator formulations usually perform significantly better than their q-operator counterparts [1-4]. This is especially true for high-speed applications where the sampling rate is much larger than the underlying system bandwidth. Under these conditions, q-operator formulated discrete-time systems tend to become ill-conditioned [1-2].

Although a large amount of work is available on the effects of coefficient sensitivity and quantization noise, a deterministic study of the nonlinear behavior of discrete-time systems formulated with the δ-operator has not been undertaken. In the case of floating-point (FLP) arithmetic, some results for feedback system are available in [2].

In this work, we focus on the convergence behavior of the unforced system response and global asymptotic stability of δ-operator formulated discrete-time systems implemented in fixed-point (FXP) arithmetic. In particular, via necessary conditions for stability, it will be shown that such systems tend to produce DC limit cycles. We will also perform a deterministic analysis of the finite wordlength properties of multi-dimensional δ-operator implemented discrete time systems. The stability behavior in the m-D case has not been previously investigated, although convergence to the true equilibrium point(s) is one of the most fundamental requirements for any discrete time system realization.

The structure of this article is as follows: In Section II, we introduce notation and nomenclature for the 1-D case. The model for 1-D δ-operator formulated discrete-time systems, with and without quantization nonlinearities, is briefly discussed. Section III addresses the problem of asymptotic stability for the 1-D case. In terms of ensuing DC limit cycles, necessary conditions for global asymptotic stability are formulated. It is shown that, when FXP arithmetic is used, stability of the linear system is often lost. Bounds on the

size of the deadbands are also provided. In section IV, the multidimensional case is investigated using sets of 1-D conditions for asymptotic stability. Section V provides concluding remarks.

II. NOTATION AND NOMENCLATURE

Since our focus is the investigation of stability properties of δ-operator formulated discrete-time systems under unforced conditions, the state equations of the system under zero-input will be considered.

In the linear case, the general m-th order state-space representation is given by

$$\delta[\mathbf{x}](n) = A^{\delta}\mathbf{x}(n); \tag{1}$$

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n), \tag{2}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_m(n)]^T$ is the state vector at instant n, $A^{\delta} = \{a_{ij}^{\delta}\} \in \Re^{m \times m}$ is the system matrix, and $\Delta > 0$ is the sampling time. Moreover, $\delta[\cdot]$ represents the δ-operator, that is,

$$\delta[x_\nu](n) = \frac{x_\nu(n+1) - x_\nu(n)}{\Delta}, \ \forall \nu = 1, \ldots, m, \tag{3}$$

and $\delta[\mathbf{x}](n) = [\delta[x_1](n), \ldots, \delta[x_m](n)]^T$. A δ -system is stable, if and only if the following condition on the eigenvalues λ_i^{δ} of the matrix A^{δ} is satisfied [1]:

$$| \lambda_i^{\delta} - \Delta^{-1} | < \Delta^{-1}, \quad i = 1, \cdots, m.$$

Therefore a stable system matrix cannot be defective, i.e. it cannot have a zero eigenvalue.

The actual implementation of (1) and (2) in FXP format gives rise to nonlinear quantization operations that occur at various locations depending on the hardware realization.

Eqn. (1) can be implemented either by using single wordlength accumulators (creating a quantization error after each multiplication) or by using double wordlength accumulators (creating a quantization error only after summation). We will only consider the latter option since practically all modern DSP machines offer double precision accumulators.

Eqn. (1) can then be written as

$$\delta[\mathbf{x}](n) = Q\{A^{\delta}\mathbf{x}(n)\}, \tag{4}$$

where Q is a vector-valued quantization nonlinearity of the form

$$Q\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}. \tag{5}$$

Here, $Q\{x_\nu\}$ can denote magnitude truncation, two's complement truncation, or rounding.

Eqn. (2) can be implemented in two different ways:

$$\mathbf{x}(n+1) = \mathbf{x}(n) + Q\{\Delta \cdot \delta[\mathbf{x}](n)\}, \tag{6}$$

or

$$\mathbf{x}(n+1) = Q\{\mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n)\}. \tag{7}$$

Eqn. (6) corresponds to quantization after multiplication while (7) corresponds to quantization after summation. In contrast to (1), for equation (2), it is not clear which of the two quantization schemes in (6) and (7) is preferable. We will therefore consider both possibilities.

Throughout this paper, we will use the following definition of stability:

Definition. The discrete-time system in (4,6) or (4,7) is globally asymptotically stable if and only if, for any initial condition $\mathbf{x}(0)$, the state vector $\mathbf{x}$ asymptotically reaches zero, that is, $\mathbf{x}(n) \to \mathbf{0}$ for $n \to \infty$.

Comment. Since the FXP systems considered are in fact finite state machines, the condition $\mathbf{x}(n) \to \mathbf{0}$ for $n \to \infty$ may be strengthened to $\mathbf{x}(N) = \mathbf{0}$ for some finite N [5].

The following additional symbols will be used:

l: quantization step size

$\underline{0}, \underline{1}$: Vector with all elements being zero or one, respectively.

$Int(x)$: the largest integer function, i.e. the largest integer smaller than or equal to x.

$\mathcal{D}_\delta^{MT}, \mathcal{D}_\delta^{R}, \mathcal{D}_\delta^{TWO}$: Deadbands in terms of the incremental difference vector for magnitude truncation, rounding and two's complement, respectively.

$\mathcal{D}_x^{MT}, \mathcal{D}_x^{R}, \mathcal{D}_x^{TWO}$: Deadbands in terms of the state vector $\mathbf{x}$ for magnitude truncation, rounding and two's complement truncation, respectively.

$\mathcal{A}_\delta^{MT}, \mathcal{A}_\delta^{R}, \mathcal{A}_\delta^{TWO}$: corresponding deadband for the unquantized difference vector.

$\mathcal{H}_L^{MT}$: largest hypercube embedded in $\mathcal{D}_x^{MT}$.

$\mathcal{H}_U^{MT}$: smallest hypercube embedding $\mathcal{D}_x^{MT}$.

III. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY

III.1 DC Limit Cycles

First, we will consider the system described by (4,6). From the definition for global asymptotic stability as stated in the previous section, it is necessary that

$$Q\{\Delta \cdot \delta[\mathbf{x}](n)\} \neq \mathbf{0}, \quad \text{for any} \quad \mathbf{x}(n) \neq \mathbf{0}. \tag{8}$$

This is just one of a finite set of conditions that is required to ensure global asymptotic stability of a FXP implementation of a linearly stable system [5].

The following theorem on global asymptotic stability of delta-operator formulated discrete time systems provides conditions on the sampling time:

Theorem 1. A necessary condition for global asymptotic stability of the δ-operator formulated discrete-time system in (4,6) is $\Delta \geq 0.5$ for rounding and $\Delta \geq 1$ for truncation.

Proof: At first, we will address the case of magnitude rounding: The necessary condition

for global asymptotic stability (8) is violated, if

$$| \Delta \cdot \delta[x_\nu](n) | < \frac{l}{2} \quad for \quad \nu = 1, \cdots, m. \tag{9}$$

and $\delta[\mathrm{x}](n) \neq \mathbf{0}$. With

$$\delta[x_\nu](n) = l \quad for \quad \nu = 1, \cdots, m, \tag{10}$$

we can rewrite (9) as

$$\Delta < \frac{1}{2}. \tag{11}$$

If the sampling time is chosen according to (11), then condition (9) is satisfied and hence, the system will exhibit a period one limit cycle. Therefore, in order to avoid a period one limit cycle we require

$$\Delta \geq \frac{1}{2} \tag{12}$$

(Additional constraints will have to be imposed on Δ in order to guarantee the absence of limit cycles with a period other than one.) This proves the Theorem for rounding.

In the case of magnitude truncation, equation(9) becomes:

$$| \Delta \cdot \delta[x_\nu](n) | < l \qquad for \quad \nu = 1, \cdots, m \tag{13}$$

with $\delta[\mathrm{x}](n) \neq \mathbf{0}$ With (13) and (10), one arrives at the following condition, which excludes period one limit cycles:

$$\Delta \geq 1 \tag{14}$$

For two's complement truncation, equation (9) takes the form:

$$0 \leq \Delta \cdot \delta[x_\nu](n) < l \tag{15}$$

Together with (10), the above equation also results in (14), which proves the Theorem.

The above theorem shows that high-speed δ-operator formulated implementations that possess a small sampling time cannot be realized limit cycle free in FXP format! Since the advantages of delta-operator systems with respect to coefficient sensitivity and quantization

noise require a short sampling time much smaller than one, this requirement cannot be met if limit cycles have to be avoided.

A second necessary condition for the system in $\{(4),(6)\}$ can be obtained by noting that

$$\delta[\mathbf{x}](n) = \mathbf{0} \tag{16}$$

can occur in (4) even though the state vector $\mathbf{x}(n) \neq \mathbf{0}$.

Therefore, for magnitude rounding, no nonzero state vector $\mathbf{x}(n)$ that belongs to the quantization lattice and satisfies

$$-\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} < A^{\delta} \cdot \mathbf{x}(n) < +\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \tag{17}$$

may be allowed to exist. In (17), the inequality has to hold elementwise.

Equation (17) has the following geometric interpretation:

Each of the resulting m inequalities can be geometrically interpreted in the state space as the intersection of two half spaces in $\Re^m$. These intersections are symmetric about the origin and have parallel boundaries. The normal vector to the boundaries is given by the particular row vector of A^{δ}. Only if the intersection of *all* such m half spaces contains at least one nonzero point in $\Re^m$ on the quantization lattice, will there exist a nonzero state vector that is an equilibrium point of the system due to equation (16). Since we only consider A^{δ} matrices, which are stable, the system matrix A^{δ} is always invertible. One can therefore rewrite (1) to obtain a sufficient condition for the existence of non-zero state vectors, which are equilibrium points due to equation (16):

$$\mathbf{x}(n) = (A^{\delta})^{-1}\delta[\mathbf{x}](n) \quad with \qquad \delta[\mathbf{x}](n) \in (-l/2, l/2)^m \tag{18}$$

In order to obtain bounds for each of the components of $\mathbf{x}(n)$ we use the infinity norm:

$$\| \mathbf{x}(n) \|_{\infty} \leq \| (A^{\delta})^{-1} \|_{\infty} \| \delta[\mathbf{x}](n) \|_{\infty} < \| (A^{\delta})^{-1} \|_{\infty} \frac{l}{2} \tag{19}$$

The perallelepiped described by (18) is therefore imbedded in the hypercuboid described by (19). If (19) does not permit any points $\mathbf{x(n)}$ of the sampling lattice, instability due to (16) cannot occur. From (19), this is the case if

$$\| (A^\delta)^{-1} \|_\infty < 2. \tag{20}$$

Eqn. (16) can also be interpreted from an eigenvalue/eigenvector viewpoint. In high-speed digital filters where the sampling frequency is typically much higher than the bandwidth of the processed signal, the eigenvalues of a q-operator implementation cluster around the point $z = 1$ [1]. The corresponding δ-operator implementation for large sampling times has eigenvalues clustered around zero. However, as the sampling time becomes small, these eigenvalues move towards the eigenvalues of the underlying continuous-time system [1]. In other words, for large sampling times, the system matrix will be ill-conditioned, that is, vectors $\mathbf{x}(n) \neq \mathbf{0}$ exist such that $A^\delta \cdot \mathbf{x}(n)$ is close to the zero vector. According to (16), this is likely to cause a DC limit cycle. For small sampling times, this problem may not occur; however, in this case, the conditions in Theorem 1 are not satisfied and the system is already known to produce limit cycles.

In the case of the remaining two quantization schemes, the inequalities corresponding to (17) are given below: For two's complement truncation,

$$\mathbf{0} \leq A^\delta \cdot \mathbf{x}(n) < \begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}, \tag{21}$$

and, for magnitude truncation,

$$-\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix} < A^\delta \cdot \mathbf{x}(n) < +\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}. \tag{22}$$

Again, the above inequalities have to interpreted elementwise. The embedding hypercubes can be constructed for the perallelepiped in (21) and (22) in a similar fashion as for rounding in (18).

So far, we only addressed the system described by (4,6). A similar analysis can be conducted for the system in (4,7). Since (4) is common to both realizations, equations

(17,21,22) are still valid and provide conditions under which the finite difference is quantized to zero and a DC limit cycle is produced. We will now briefly discuss necessary conditions for global asymptotic stability obtained from (7).

A period one limit cycle exists, if the condition

$$\mathbf{x} = Q(\mathbf{x} + \Delta\delta[\mathbf{x}](n)) \tag{23}$$

is satisfied for $\mathbf{x} \neq \mathbf{0}$. Using a similar argument as in the proof of Theorem 1, for rounding, equation (23) is satisfied if:

$$-\frac{l}{2} \leq \Delta\delta[x_\nu](n) < \frac{l}{2} \qquad for \quad x_\nu > 0 \tag{24}$$

$$-\frac{l}{2} < \Delta\delta[x_\nu](n) \leq \frac{l}{2} \qquad for \quad x_\nu < 0 \tag{25}$$

$$-\frac{l}{2} < \Delta\delta[x_\nu](n) < \frac{l}{2} \qquad for \quad x_\nu = 0 \tag{26}$$

$$\nu = 1, \cdots, m$$

Therefore

$$\Delta > \frac{1}{2} \tag{27}$$

is required to exclude period one limit cycles.

For magnitude truncation, (23) is satisfied, if

$$0 \leq \Delta\delta[x_\nu](n) < l \qquad for \quad x_\nu > 0 \tag{28}$$

$$-l < \Delta\delta[x_\nu](n) \leq 0 \qquad for \quad x_\nu < 0 \tag{29}$$

$$-l < \Delta\delta[x_\nu](n) < l \qquad for \quad x_\nu = 0 \tag{30}$$

$$\nu = 1, \cdots, m$$

In the case of two's complement truncation, the condition for a DC limit cycle is simply given by

$$0 \leq \Delta\delta[x_\nu](n) < l, \qquad \nu = 1, \cdots, m. \tag{31}$$

The conditions (28-30) and (31) again result in the condition $\Delta \geq 1$ for the absence of period one limit cycles.

We therefore obtain almost the same conclusion as for the previously considered system:

$$\Delta > \frac{1}{2} \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1 \quad \text{for truncation.}$$

Therefore, Theorem 1 also holds for the system representation in $\{(4),(7)\}$, if the condition for rounding is slightly changed to $\Delta > \frac{1}{2}$.

Upto now, we provided necessary conditions for stability of delta-operator fomulated discrete time systems in fixed point arithmetic. Since it has been established, that for small sampling periods, the delta-operator systems always exhibits period one limit cycles, one needs to examine the amplitude of these limit cycles for a given sampling time in order to obtain further insight into the practical impact of this problem. In what follows, bounds on the deadbands will be derived as a function of the A^δ-matrix and the sampling time Δ.

III.2 Deadband Bounds

This subsection provides an answer to the question of the size of the limit cycle amplitudes. Given a sampling time Δ and a system matrix A^δ, bounds for the deadbands as well as the deadband geometry will be described. This will be done in detail for the case of magnitude truncation. For magnitude rounding and two's complement truncation, the results will be stated briefly without proof. Since the results for the system (4,7) are very similar to the results for the system(4,6), this subsection focuses only on the latter.

For each quantization scheme, we will provide the geometry of the deadband in terms of the incremental difference vector as well as the state vector. Two hypercubes, which bound the deadband region from the inside and the outside are also derived for each case.

Theorem 2:

For the system (4,6) implemented in magnitude truncation, the deadband (in terms of

period one limit cycles) in the incremental difference vector space is given by:

$$\mathcal{D}_\delta^{MT} = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty \leq [Int(\Delta^{-1}) - 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{32}$$

and

$$\mathcal{D}_\delta^{MT} = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty \leq Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1}. \tag{33}$$

The corresponding period one limit cycle deadband in the state space is given by

$$\mathcal{D}_x^{MT} = \{\mathbf{x} \mid \mathbf{x} = (A^\delta)^{-1}\delta[\mathbf{x}], \quad \delta[\mathbf{x}] \in \mathcal{A}_\delta^{MT}\} \tag{34}$$

where

$$\mathcal{A}_\delta^{MT} = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < [Int(\Delta^{-1}) + 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1} \tag{35}$$

and

$$\mathcal{A}_\delta^{MT} = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{36}$$

Proof:

The proof will be carried out for $Int(\Delta^{-1}) \neq \Delta^{-1}$, since the case $Int(\Delta^{-1}) = \Delta^{-1}$ follows in a similar fashion. From (13), the expression for period one limit cycles can be expressed as

$$\| \Delta\delta[\mathbf{x}](n) \|_\infty < l. \tag{37}$$

Solving (37) for $\delta[\mathbf{x}]$ and considering, that $\delta[\mathbf{x}]$ produced by equation (4) is an integer multiple of the quantization step l, one obtains

$$\| \delta[\mathbf{x}](n) \|_\infty \leq Int(\Delta^{-1}) \cdot l \tag{38}$$

for $Int(\Delta^{-1}) \neq \Delta^{-1}$ which is the hypercube in (33). Now consider the following slightly larger hypercube $\mathcal{A}_\delta^{MT}$ in $\delta[\mathbf{x}]$:

$$\mathcal{A}_\delta^{MT} = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < [Int(\Delta^{-1}) + 1]l\} \tag{39}$$

$\mathcal{A}_\delta^{MT}$ describes the open set of all incremental difference vectors, which, after quantization will be mapped into the hypercube $\mathcal{D}_\delta^{MT}$, i.e.

$$Q(\delta[\mathbf{x}]) \in \mathcal{D}_\delta^{MT}, \quad \forall \delta[\mathbf{x}] \in \mathcal{A}_\delta^{MT}.$$

Therefore the deadband in terms of $\mathbf{x}$ can simply be found by determining the set of all $\mathbf{x}$, which satisfy

$$A^{\delta}\mathbf{x} \in \mathcal{A}_{\delta}^{MT}.$$

Since A^{δ} was assumed to be linearly stable, it is also invertible . Therefore the deadband in the state space is obtained by

$$\mathcal{D}_{x}^{MT} = \{\mathbf{x} \mid \mathbf{x} = (A^{\delta})^{-1}\delta[\mathbf{x}], \quad \delta[\mathbf{x}] \in \mathcal{A}_{\delta}^{MT}\}$$

This completes the proof for $Int(\Delta^{-1}) \neq \Delta^{-1}$.

The following Corollary provides the largest hypercube in the state space, which is contained in the perallelepiped $\mathcal{D}_{x}^{MT}$. This result allows to obtain the largest magnitude of state vector components, which can still belong to the deadband. It also provides a simple upper bound on the volume of the deadband.

Corollary 3:

The largest hypercube $\mathcal{H}_{L}^{MT}$ embedded in $\mathcal{D}_{x}^{MT}$ is given by:

$$\mathcal{H}_{L}^{MT} = \{\mathbf{x} \mid \quad \| \mathbf{x}(n) \|_{\infty} < \frac{[Int(\Delta^{-1}) + 1]l}{\| A^{\delta} \|_{\infty}}\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1} \tag{40}$$

and by

$$\mathcal{H}_{L}^{MT} = \{\mathbf{x} \mid \quad \| \mathbf{x}(n) \|_{\infty} < \frac{Int(\Delta^{-1})l}{\| A^{\delta} \|_{\infty}}\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{41}$$

Proof:

Assume $Int(\Delta^{-1}) \neq \Delta^{-1}$. From (1) we obtain for the unquantized incremental difference vector:

$$\| \delta[\mathbf{x}] \|_{\infty} \leq \| A^{\delta} \|_{\infty} \| \mathbf{x} \|_{\infty} \tag{42}$$

Since $\mathcal{A}_{\delta}^{MT}$ describes the set of unquantized difference vectors, which after quantization maps into the deadband region $\mathcal{D}_{\delta}^{MT}$, one can use the right side of (42) to ensure, that equation (39) is satisfied and obtain:

$$\| A^{\delta} \|_{\infty} \| \mathbf{x} \|_{\infty} < [Int(\Delta^{-1}) + 1] \cdot l \tag{43}$$

Solving (43) for $\| \mathbf{x} \|_\infty$ produces the desired result. Since $\mathcal{H}_L^{MT}$ is a hypercube centered at the origin, there exists a $\mathbf{x} \in \mathcal{H}_L^{MT}$, such that

$$\| \delta[\mathbf{x}] \|_\infty = \| A^\delta \|_\infty \cdot \| \mathbf{x} \|_\infty \, . \tag{44}$$

Hence this is the largest such hypercube. The proof for the case $Int(\Delta^{-1}) = \Delta^{-1}$ follows from (43) in a similar fashion.

The next Corollary provides the smallest hypercube in the state space, which still contains $\mathcal{D}_x^{MT}$. This provides a lower bound on the volume of the deadband:

Corollary 4:

The smallest hypercube $\mathcal{H}_U^{MT}$ containing $\mathcal{D}_x^{MT}$ is given by

$$\mathcal{H}_U^{MT} = \{ \mathbf{x} \mid \quad \| \mathbf{x} \|_\infty < \| (A^\delta)^{-1} \|_\infty \, (Int(\Delta^{-1}) + 1) \cdot l \} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1} \tag{45}$$

and

$$\mathcal{H}_U^{MT} = \{ \mathbf{x} \mid \quad \| \mathbf{x} \|_\infty < \| (A^\delta)^{-1} \|_\infty \, Int(\Delta^{-1}) \cdot l \} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{46}$$

Proof:

At first consider the case $Int(\Delta^{-1}) \neq \Delta^{-1}$: From (1) we have for the unquantized state vector:

$$\mathbf{x} = (A^\delta)^{-1} \delta[\mathbf{x}] \tag{47}$$

Taking norms and using the inequality in (35), we obtain the following open hypercube, which contains $\mathcal{D}_x^{MT}$:

$$\| \mathbf{x} \|_\infty \leq \| (A^\delta)^{-1} \|_\infty \| \delta[\mathbf{x}] \|_\infty < \| (A^\delta)^{-1} \|_\infty \, [Int(\Delta^{-1}) + 1] \cdot l$$

Since $\mathcal{D}_\delta^{MT}$ is a hypercube centered at the origin, there exists a $\delta[\mathbf{x}]$, such that

$$\| (A^\delta)^{-1} \|_\infty \| \delta[\mathbf{x}] \|_\infty = \| \mathbf{x} \|_\infty \, .$$

Hence $\mathcal{H}_U^{MT}$ is the smallest such hypercube. The proof for the case $Int(\Delta^{-1}) = \Delta^{-1}$ is identical and requires the use of (36) instead of (35).

Remarks:

1. Since $\| (A^\delta)^{-1} \| \cdot \| (A^\delta) \| \geq 1$, we have $\mathcal{H}_L^{MT} \subset \mathcal{H}_U^{MT}$. For matrices which satisfy

$$\| (A^\delta)^{-1} \|_\infty \cdot \| A^\delta \|_\infty = 1 \tag{48}$$

the two hypercubes are identical and coincide with the deadband region $\mathcal{D}_x^{MT}$, i.e. $\mathcal{H}_U^{MT} = \mathcal{H}_L^{MT} = \mathcal{D}_x^{MT}$.

2. $\mathcal{D}_\delta^{MT}$ is a closed m-D hypercube, centered around the origin. Its boundary coincides with points of the quantization lattice. The faces of the hypercube are orthogonal to the corresponding axis of the incremental difference vector space.

3. $\mathcal{D}_x^{MT}$ is an open parallelepiped. $\mathcal{D}_x^{MT}$ describes the deadband for period one limit cycles in terms of the state vector.

4. The total deadband includes the region $\mathcal{D}_\delta^{MT}$ ($\mathcal{D}_x^{MT}$) for the incremental difference vector (the state vector.)

5. A useful measure of the deadband size in terms of the state vector $\mathbf{x}$ is the volume in the state space. The 'volume' Vol_δ of the deadband in $\delta[\mathbf{x}]$ is easily computable due to the hypercube geometry. From (47) we obtain for the volume Vol_x in the state space of $\mathbf{x}$:

$$Vol_x = det((A^\delta)^{-1}) \cdot Vol_\delta \tag{49}$$

6. Given a realization, increasing the sampling rate (Δ^{-1}) will result in a larger deadband.

The relationships for the deadband of quantization schemes other than magnitude truncation are given below:

Magnitude Rounding:

$$\mathcal{D}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty \leq [Int(\frac{1}{2}\Delta^{-1}) - 1] \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) = \frac{1}{2}\Delta^{-1} \tag{50}$$

and

$$\mathcal{D}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty \leq Int(\frac{1}{2}\Delta^{-1}) \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) \neq \frac{1}{2}\Delta^{-1} \tag{51}$$

For the deadband in terms of the state vector we have:

$$\mathcal{D}_x^R = \{\mathbf{x} \mid \mathbf{x} = (A^\delta)^{-1}\delta[\mathbf{x}], \delta[\mathbf{x}] \in \mathcal{A}_\delta^R\} \tag{52}$$

where

$$\mathcal{A}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < [Int(\frac{1}{2}\Delta^{-1}) - \frac{1}{2}] \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) = \frac{1}{2}\Delta^{-1} \tag{53}$$

and

$$\mathcal{A}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < [Int(\frac{1}{2}\Delta^{-1}) + \frac{1}{2}] \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) \neq \frac{1}{2}\Delta^{-1} \tag{54}$$

Two's Complement Truncation:

$$\mathcal{D}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] \leq \underline{1} \cdot [Int(\Delta^{-1}) - 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{55}$$

and

$$\mathcal{D}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] \leq \underline{1} \cdot Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1}. \tag{56}$$

For the deadband in terms of the state vector we have:

$$\mathcal{D}_x^{TWO} = \{\mathbf{x} \mid \mathbf{x} = (A^\delta)^{-1}\delta[\mathbf{x}], \quad \delta[\mathbf{x}] \in \mathcal{A}_\delta^{TWO}\} \tag{57}$$

where

$$\mathcal{A}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] < \underline{1} \cdot Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{58}$$

and

$$\mathcal{A}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] < \underline{1} \cdot [Int(\Delta^{-1}) + 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1}. \tag{59}$$

In the above set definitions, all inequalities are to be interpreted elementwise, i.e. $\mathbf{x} \leq \mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathcal{R}^m$ means $x_i \leq y_i, \quad i = 1, \cdots, m$. Furthermore, the notation $\underline{0}, \underline{1}$ stands for the zero vector and the vector with component values of one, respectively.

IV. THE M-D CASE

IV.1 Additional Notation for the m-D Case

The m-D Roesser model has the following δ-operator formulation [7]:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \begin{bmatrix} B_1^{\delta} \\ \vdots \\ B_m^{\delta} \end{bmatrix} \mathbf{u}(\mathbf{n}); \tag{60}$$

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix}. \tag{61}$$

The input-state equations in (60) and (61) describe a first hyper-quadrant causal m-D system with a uniform sampling period of Δ in all directions. The operators $q^{(i)}$ and $\delta^{(i)}$ represent the shift- and delta-operator in the direction specified by the axis n_i. In particular

$$q^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) \tag{62}$$

$$\delta^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \frac{1}{\Delta}(\mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) - \mathbf{x}^{(i)}(\mathbf{n})). \tag{63}$$

Here, $(\mathbf{n}) \doteq (n_1, \ldots, n_m)$ denotes a point in the first hyper-quandrant, $\mathbf{x}^{(i)}(\mathbf{n})$ is the portion of the state vector propagating in the direction specified by the axis n_i, $u(\mathbf{n})$ is the m-D input vector, and A_{ij}^{δ} and B_i^{δ}, for $i = 1, \ldots, m$, $j = 1, \ldots, m$, are the submatrices of the system and input matrices, respectively.

If (60) is realized in fixed-point arithmetic, it takes the following form under zero-input conditions:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} \right\} \tag{64}$$

Equation (64) assumes quantization after summation; since practically all modern DSP machines implement this quantization scheme, we only consider this format. The

vector-valued quantization nonlinearity $\mathbf{Q}\{\cdot\}$ may represent any one of the conventional schemes, viz., magnitude truncation, magnitude rounding, two's complement truncation, and two's complement rounding.

Equation (61) can be implemented in two different forms:

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(i)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \tag{65}$$

or

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\}. \tag{66}$$

Equation (65) corresponds to quantization after multiplication, whereas (66) corresponds to quantization after addition. In contrast to (60), for (61), it is not obvious which of the two forms stated above is preferable.

The following definition for asymptotic stability [8] will be used throughout this paper.

Definition. An m-D first hyper-quadrant causal discrete-time system is asymptotically stable under all finitely extended bounded input signals $u(\mathbf{n})$ where

$$|u(\mathbf{n})| \leq S, \quad \text{for} \quad n_1 + \cdots + n_m \leq D; \tag{67}$$

$$u(\mathbf{n}) = 0, \quad \text{for} \quad n_1 + \cdots + n_m > D, \tag{68}$$

if all the states of the m-D discrete-time system asymptotically reach zero for $n_1 + \cdots + n_m \to \infty$. Here, $n_\nu \geq 0$, $\nu = 1, \ldots, m$, S is a nonnegative real number, and D is a positive integer.

Since the fixed-point systems considered are in fact finite state machines, the condition

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} \to \mathbf{0},$$

for $n_1 + \cdots + n_m \to \infty$, $n_\nu \geq 0$, $\nu = 1, \ldots, m$, can be strengthened to

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} = \mathbf{0},$$

for all points $n_1 + \cdots + n_m \geq c$, $n_\nu \geq 0$, $\nu = 1, \ldots, m$, where c is some finite integer.

IV.2 Necessary Conditions for Global Asymptotic Stability

In this section, we present some necessary conditions for stability of a first hyper-quadrant causal m-D discrete-time system represented in its Roesser local state-space model in (60,61). These necessary conditions are formulated in terms of 1-D conditions. This theorem follows directly from a result in [6] which was formulated for q-operator implemented discrete-time systems. The proof of the theorem rests on the fact that a first hyper-quadrant m-D system can be described by a 1-D system for those locations that are along the m coordinate axes of the boundary of the hyper-quadrant. Reformulating the result in [6] for δ-operator systems produces the following theorem:

Theorem 5.

(a) A necessary condition for global asymptotic stability of the system in (64,65) is that each of the following 1-D systems in (69,70) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{69}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{x}^{(i)}(n_i) + \mathbf{Q}\left\{\Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\}, \tag{70}$$

where $i = 1, \ldots, m$.

(b) A necessary condition for global asymptotic stability of the system in (64,66) is that each of the following in 1-D systems in (71,72) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{71}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{\mathbf{x}^{(i)}(n_i) + \Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\}, \tag{72}$$

where $i = 1, \ldots, m$.

Proof. For a detailed proof, and generalizations to higher sub-dimensional systems, the reader is referred to [6].

Theorem 5 can be viewed as an extension of the concept of practical BIBO stability to asymptotic stability of nonlinear systems. It is particularly useful in proving instability in m-D nonlinear systems.

We can now combine Theorem 1 and Theorem 5 to formulate a necessary condition for stability of m-D first hyper-quadrant causal δ-operator formulations of the generalized Roesser model.

Corollary 6.

(a) A necessary condition for global asymptotic stability of the m-D systems in (64,65) is

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1, \quad \text{for truncation.}$$

(b) A necessary condition for global asymptotic stability of the $m-$D system in (64,66) is

$$\Delta > 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1, \quad \text{for truncation.}$$

Proof. The proof follows from Theorems 1 and 5.

Remarks:

1. Corollary 6 is also essentially applicable to the case where the sampling time varies with the direction of propagation. In the case of the system description (64,65), the inequalities in Corollary 6 would have to be replaced by

$$\Delta_i \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta_i \geq 1, \quad \text{for truncating,}$$

for $i = 1, \ldots, m$. The conditions for the system (64,66) are analogous.

2. Our analysis is limited to the zero-input case for which DC limit cycles along the axis were used to derive conditions for non-convergence. If one includes other types of limit cycles in the analysis or even response types, which are not periodic and are known to exist only in the m-D case, the requirements for Δ may become even more severe.

3. Corollary 6 shows that fixed-point implementations of 1-D and m-D δ-operator systems *cannot be realized limit cycle free, if good coefficient sensitivity and quantization noise measures have to be achieved.*

V. CONCLUSION

In this paper, it was shown that fixed-point implementations of 1-D and m-D δ-operator systems are not limit cycle free even if the underlying linear system is stable and the sampling time is chosen small. This non-convergent behavior can be explained by the quantization of the δ-term to zero which leaves the state vector unchanged. The smaller the sampling time, the more severe this effect. The size of the deadband increases with a decreasing sampling time. Therefore, the practical value of δ-operators for fixed-point implementations of 1-D and m-D systems is questionable. There are however indications that this effect is much less severe in floating-point implementations.

δ-operator implemented discrete-time systems represent a class of systems where the quantization noise at the output can be small compared to other realizations. However, as was shown above, such realizations will invariably exhibit limit cycles, which are highly correlated quantization noise. Therefore, in this case, typical measures for quantization noise are of very limited use for obtaining any insight into the likelihood of limit cycles and vice versa.

ACKNOWLEDGEMENT

This work was supported by two grants from the Office of Naval Research (ONR): N 00014-94-1-0454 and N 00014-94-1-0387.

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE,* vol. 80, no. 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta operators," *IEEE Transactions on Automatic Control,* vol. 31, pp. 1015-1021, Nov. 1986.

[3] G.Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the IEEE Conference on Decision and Control (CDC'90),* vol. 2, pp. 954-959, Honolulu, HI, 1990.

[4] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations", *IEEE Transactions on Signal Processing*, Vol. 41, No. 2, pp. 629-637, Feb. 1993.

[5] P. H. Bauer and L. J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed point digital filters", *IEEE Transactions on Signal Processing*, Vol. 39, No. 11, pp. 2400-2410, Nov. 1991.

[6] P. Bauer, "Low-dimensional conditions for global asymptotic stability of m-D nonlinear digital filters," *1994 IEEE International Symposium on Circuits and Systems,* London, England, pp. 2.461-2.464.

[7] K. Premaratne, J. Suarez, M. Ekanayake, P.H. Bauer, "Two-dimensional delta-operator formulated discrete time systems: State space realization and its coefficient sensitivity properties", *Proceedings of the 37th Midwest Symposium on Circuits and Systems*, Aug. 1994, Lafayette, LA

[8] P. Bauer, "Finite wordlength effects in $m-$D digital filters with singularities on the stability boundary," *IEEE Transactions on Signal Processing* vol. 40, no. 4, pp. 894-900, April 1994.

ZERO-CONVERGENCE OF 2-D ROESSER STATE SPACE MODELS IMPLEMENTED IN FLOATING POINT ARITHMETIC

Peter H. Bauer
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
Tel: (219) 631-8015
e-mail: pbauer@mars.ee.nd.edu

Kamal Premaratne
Department of Electrical & Computer Engineering
University of Miami
Coral Gables, FL 33124
Tel: (305) 284-4051
e-mail: kprema@umiami.ir.miami.edu

Abstract:

Zero input asymptotic response behavior of general order 2-D digital filters with floating point arithmetic is investigated. In particular, conditions for the absence of so-called R1 and R2 responses (large amplitude limit cycles) are provided for 2-D first quarter-plane causal filters.

I. Introduction

Recently, floating point arithmetic has become popular for a number of digital signal processing applications. The implementation of digital filters in floating point format is especially attractive due to the high dynamical range and the high-level programming tools available.

Previous work on the convergence behavior of floating point digital filters concentrated on 1-D second order system [1]. Some results on direct form filters are also available [2]. However, to the authors' knowledge, the case of general order 1-D or 2-D state space models has not been tackled.

This paper provides such an analysis which can be applied to any digital filter structure of arbitrary order and dimension one or two. In order to avoid distinguishing among a number of reformatting and quantization schemes, the result introduced in this paper takes a parameterization approach to the error description. This allows to apply the derived result to any type of floating point format.

II. Preliminaries

Consider the Roesser model for the first quarter-plane causal 2-D system:

$$\begin{pmatrix} \underline{\hat{x}}^h(n_1+1,n_2) \\ \underline{\hat{x}}^v(n_1,n_2+1) \end{pmatrix} = \begin{pmatrix} A_{hh} & A_{hv} \\ A_{vh} & A_{vv} \end{pmatrix} \begin{pmatrix} \underline{\hat{x}}^h(n_1,n_2) \\ \underline{\hat{x}}^v(n_1,n_2) \end{pmatrix} \tag{1}$$

$$\begin{aligned} A &= \begin{pmatrix} A_{hh} & A_{hv} \\ A_{vh} & A_{vv} \end{pmatrix}, \; A \in \Re^{(N_1+N_2)\times(N_1+N_2)} \\ & A_{hh} \in \Re^{N_1 \times N_1} \\ & A_{vv} \in \Re^{N_2 \times N_2} \end{aligned} \tag{2}$$

The submatrices A_{vh} and A_{hv} are of the appropriate dimensions. The vectors $\underline{\hat{x}}^h$ and $\underline{\hat{x}}^v$ are horizontally and vertically propagating state vectors of the ideal system, respectively.

For floating point realizations of (1), the following error model describes the system behavior:

$$\begin{pmatrix} \underline{x}^h(n_1+1,n_2) \\ \underline{x}^v(n_1,n_2+1) \end{pmatrix} = \begin{pmatrix} A_{hh} & A_{hv} \\ A_{vh} & A_{vv} \end{pmatrix} \begin{pmatrix} \underline{x}^h(n_1,n_2) \\ \underline{x}^v(n_1,n_2) \end{pmatrix} + \begin{pmatrix} \underline{e}^h(n_1,n_2) \\ \underline{e}^v(n_1,n_2) \end{pmatrix} \tag{3}$$

where $\underline{e}^h(n_1,n_2) \in \Re^{N_1}$, $\underline{e}^v(n_1,n_2) \in \Re^{N_2}$ are the error vectors for the horizontally and vertically propagating states, respectively.

We also need to define the following transfer matrices:

$$\begin{pmatrix} \underline{X}^h(z_1,z_2) \\ \underline{X}^v(z_1,z_2) \end{pmatrix} = \begin{pmatrix} H^{hh}(z_1,z_2) & H^{hv}(z_1,z_2) \\ H^{vh}(z_1,z_2) & H^{vv}(z_1,z_2) \end{pmatrix} \begin{pmatrix} \underline{E}^h(z_1,z_2) \\ \underline{E}^v(z_1,z_2) \end{pmatrix} \tag{4}$$

where

$$\begin{pmatrix} H^{hh}(z_1,z_2) & H^{hv}(z_1,z_2) \\ H^{vh}(z_1,z_2) & H^{vv}(z_1,z_2) \end{pmatrix} = \left(\begin{bmatrix} z_1 I_1 & \phi \\ \phi & z_2 I_2 \end{bmatrix} - A \right)^{-1} \tag{5}$$

In Equation (4), $\underline{X}^h(z_1,z_2)$ and $\underline{X}^v(z_1,z_2)$ are the z-transforms of the states $\underline{x}^h(n_1,n_2)$ and $\underline{x}^v(n_1,n_2)$, respectively. The transforms $H^{hh}(z_1,z_2), H^{hv}(z_1,z_2), H^{vh}(z_1,z_2)$ and $H^{vv}(z_1,z_2)$ are transfer submatrices of dimensions $N_1 \times N_1, N_1 \times N_2, N_2 \times N_1, N_2 \times N_2$, respectively. $\underline{E}^h(z_1,z_2)$ and $\underline{E}^v(z_1,z_2)$ are the 2-D z-transforms of the error signal vectors $\underline{e}^h(n_1,n_2)$ and $\underline{e}^v(n_1,n_2)$, respectively. Furthermore, in (5), I_1 and I_2 denote identity matrices of dimensions $N_1 \times N_1$ and $N_2 \times N_2$, respectively.

The components $H^{hh}(z_1,z_2), H^{hv}(z_1,z_2), H^{vh}(z_1,z_2)$ and $H^{vv}(z_1,z_2)$ are 2-D transforms denoted by $H_{ij}^{hh}(z_1,z_2), H_{ij}^{hv}(z_1,z_2), H_{ij}^{vh}(z_1,z_2)$ and $H_{ij}^{vv}(z_1,z_2)$, respectively.
Denoting $\mathcal{Z}\{\cdot\}$ as the 2-D z-transform, we define the following impulse responses:

$$H_{ij}^{hh}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{hh}(n_1,n_2)\};\ i=1,\cdots,N_1;\quad j=1,\cdots,N_1. \tag{6}$$

$$H_{ij}^{hv}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{hv}(n_1,n_2)\};\ i=1,\cdots,N_1;\quad j=1,\cdots,N_2. \tag{7}$$

$$H_{ij}^{vh}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{vh}(n_1,n_2)\};\ i=1,\cdots,N_2;\quad j=1,\cdots,N_1. \tag{8}$$

$$H_{ij}^{vv}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{vv}(n_1,n_2)\};\ i=1,\cdots,N_2;\quad j=1,\cdots,N_2. \tag{9}$$

Next, we define the l_1-measures for each component of the transfer function submatrices:

$$\tilde{H}_{ij}^{vv} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{vv}(n_1,n_2)|;\quad i,j=1,\cdots,N_1. \tag{10}$$

$$\tilde{H}_{ij}^{hv} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{hv}(n_1,n_2)|;\quad i=1,\cdots,N_1;\quad j=1,\cdots,N_2. \tag{11}$$

$$\tilde{H}_{ij}^{vh} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{vh}(n_1,n_2)|;\quad i=1,\cdots,N_2;\quad j=1,\cdots,N_1. \tag{12}$$

$$\tilde{H}_{ij}^{vv} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{vv}(n_1,n_2)|;\quad i,j=1,\cdots,N_2. \tag{13}$$

Also:

$$\tilde{H}_i^h = \sum_{\nu=1}^{N_1} \tilde{H}_{i\nu}^{hh} + \sum_{\nu=1}^{N_2} \tilde{H}_{i\nu}^{hv} \tag{14}$$

$$\tilde{H}_j^v = \sum_{\nu=1}^{N_1} \tilde{H}_{j\nu}^{vh} + \sum_{\nu=1}^{N_2} \tilde{H}_{j\nu}^{vv} \tag{15}$$

From [1,2] it is known that the following four state response types are encountered in floating point digital filters under zero input, if the linear filter is stable:

R1: an unbounded state response, eventually leading to overflow conditions.
R2: a bounded state response
R3: a bounded state response in underflow
R4: a zero-convergent response

III. The Main Result

The following theorem can now be formulated:

Theorem: A floating point implementation of the system in (1) for any finitely extended input signal and/or non-zero finitely extended initial conditions will produce a response type R3, if the mantissa length l_m satisfies

$$l_m \geq 2 + \log_2 \tilde{H} + \log_2 C \tag{16}$$

where $\tilde{H} = \max_{i,j}(\tilde{H}_i^h, \tilde{H}_j^v)$ and C is an implementation dependent constant.

Proof: The proof is rather lengthy and will be supplied in the final version of the paper.

Formally, this result is similar to previous results on direct forms [1] and second order state-space systems [2]. In this case, the stability margin enters the inequality through $\tilde{H}$, which is a somewhat complicated measure of the degree of stability of the system. For an unstable system $\tilde{H} \to \infty$, and for any stable system we have $\tilde{H} < \infty$. The constant C relates the magnitude of the state-variables to the error bound. This number is usually small and is directly affected by the entries of the A-matrix and the floating point format.

IV. Conclusion

This paper presents a condition on the mantissa length of a 2-D floating point digital filter of arbitrary order, which ensures convergence of the state-response into underflow, independent of the initial conditions. The mantissa length is linked to the margin of stability of the linear system as measured by $\tilde{H}$. It is also dependent on the realization itself. It should be noted that the response types R2 and R3 in the 2-D (and m-D) case do not

need to be periodic [3].

Acknowledement

This work was supported by the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N000-94-1-0387.

References

[1] P. H. Bauer, "Absolute Response Error Bounds for Floating Point Digital Filters in State Space Representation", *IEEE International Symposium on Circuits and Systems*, Chicago, May 3-6, 1993, pp. 619-622.

[2] P. H. Bauer and J. Wang, "Limit Cycle Bounds for Floating Point Implementations of Second Order Recursive Digital Filters", *IEEE Trans. on Circuits and Systems - Part II: Analog and Digital Signal Processing*, Vol. 40, No. 8, pp. 493-501, Aug. 1993.

[3] P.H. Bauer and E. I. Jury, "Nonperiodic Modes in 2-D Recursive Digital Filters under Finite Wordlength Effects", *IEEE Trans. on Circ. and Syst.*, Vol. 36, No. 7, pp. 11032-1035, 1989.

Digital Simulation of Nonlinear Systems Using Delta-Operator Based Numerical Schemes

KAMAL PREMARATNE, Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124 USA,
Tel: +1(305)284 4051; Fax: +1(305)284 4044; email: kprema@umiami.ir.miami.edu, and
PETER H. BAUER, Laboratory for Signal and Image Analysis (LISA), Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA,
Tel: +1(219)631 8015; Fax: +1(219)631 4393; email: pbauer@mars.ee.nd.edu.

Extended Abstract

This extended abstract is being submitted for possible presentation at the *IASTED International Conference on Modelling and Simulation,* Colombo, Sri Lanka, July 26-28, 1995.

Introduction

Traditional control and signal processing algorithms based on shift-operator (or, q-operator) are ill-conditioned in high performance applications that involve fast sampling/shorter wordlength [1]. In these situations, q-operator based discrete-time implementations (or, q-systems) are extremely sensitive to uncertainties inherent in modelling and parameter representation (in particular, with shorter wordlength).

Use of incremental difference operator or delta-operator (or, δ-operator) can provide an effective solution to such difficulties [1]. Compared to q-systems, δ-operator based implementations (or, δ-systems) can provide superior performance with respect to (a) coefficient sensitivity of frequency response [1], and (b) quantization noise propagation [2]. Due mainly to these, and also due to the possibility of a unified treatment of both continuous- and discrete-time systems, work on δ-systems has recently attracted considerable attention (see [1-5], and references therein).

Problem Statement

Since δ-operator can offer several important advantages over q-operator for linear, time-invariant one-dimensional (1-D) systems, would similar advantages hold true for more general classes of systems? Work on *linear,* multi-dimensional (m-D) systems indicate that this may indeed be the case [5]. In this paper, we investigate the applicability of δ-operator based numerical schemes for simulation of *nonlinear* systems.

Delta-Operator Based Numerical Scheme

q-Operator Based Numerical Scheme. We consider the computation of solution orbit of a nonlinear system of the type

$$q[\mathbf{x}](n) = \mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q), \tag{1}$$

where $q[\mathbf{x}](n) = \mathbf{x}(n+1)$. Here, $\mathbf{x}(n)$ is the state orbit $\mathbf{x} \in \Re^m$ at instant n and

Kamal Premaratne and Peter H. Bauer gratefully acknowledge the support provided by the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

$\mathbf{a}_q = [a_{1_q}, \ldots, a_{M_q}]^T \in \Re^M$ refer to system parameters that are *actually stored* within the computer while performing the iteration.

δ-Operator Based Numerical Scheme. The proposed δ-operator based scheme of the same nonlinear system in (1) is

$$\begin{aligned} \delta[\mathbf{x}](n) &= \mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) \quad \text{(Intermediate equation)} \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n) \quad \text{(Update equation)}, \end{aligned} \tag{2}$$

where $\delta[\mathbf{x}](n) = (q[\mathbf{x}](n) - \mathbf{x}(n))/\Delta$ and $\mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) = (\mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q) - \mathbf{x}(n))/\Delta$. Here, Δ is an arbitrary positive real parameter (usually the grid size) and $\mathbf{a}_\delta = [a_{\delta_1}, \ldots, a_{\delta_M}]^T \in \Re^M$ again refer to system parameters that are *actually stored* within the computer.

Now, which of the schemes (1) or (2) yield superior coefficient sensitivity of its orbit with respect to perturbation of $\mathbf{a}_q$ or $\mathbf{a}_\delta$, respectively? This consideration is crucial in high performance, real-time applications that may require fast sampling/shorter wordlength. Of course, with infinite wordlength, both (1) and (2) yield identical results. In our development, the nonlinearity is taken to belong to $\mathcal{C}^1$, that is, it possesses first partial derivatives. Small perturbations are assumed.

CONTRIBUTIONS

The contributions of this paper are the following:

1. Development of coefficient sensitivity measures M_{FXP} and M_{FLP} for fixed-point (FXP) and floating-point (FLP), repectively. These take into account that in FXP, coefficient perturbation is approximately independent of its nominal value, while in FLP, it is approximately proportional.
2. FXP: M_{FXP} for δ-system is Δ times M_{FXP} for q-system. Hence, δ-system is superior under small grid size.
3. FLP: M_{FLP} for δ-system is superior than M_{FLP} for q-system if $|a_{i_q} - 1| \leq |a_{i_q}|$, $\forall i = 1, \ldots, M$. Here, a_{i_q} indicates the 'linear' term in the i-th equation of $\mathbf{f}_q$. We show that, typical digital equivalents of continuous-time nonlinear systems obtained under fast sampling routinely satsify this condition.
4. Similar comments hold true for linear systems, piecewise $\mathcal{C}^1$ nonlinear systems, and piecewise linear systems.

REFERENCES

[1] Goodwin, G.C., Middleton, R.H., and Poor, H.V. (1992). High speed digital signal processing and control. *Proc. IEEE,* **40**, 240-259.

[2] Li, G., and Gevers, M. (1993). Roundoff noise minimization using delta operator realizations. *IEEE Trans. Sig. Proc.,* **41**, 629-637.

[3] Premaratne, K., and Bauer, P.H. (1994). Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic. *Proc. IEEE Symp. Circ. Syst. (ISCAS'94),* London, UK, **2**, 461-464.

[4] Premaratne, K., and Jury, E.I. (1994). Tabular method for determining root distribution of delta-operator formulated real polynomials. *IEEE Trans. Auto. Cont.,* **39**, 352-355.

[5] Premaratne, K., Suarez, J., Ekanayake, M.M., and Bauer, P.H. (1994). Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties. *Proc. Midwest Symp. Circ. Syst. (MWSCS'94),* Lafayette, LA; *IEEE Trans. Sig. Proc.* in review.

SouthCon '95

On Balanced Realizations of 2-D Delta-Operator Formulated Discrete-Time Systems

K. Premaratne and M.M. Ekanayake
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124 USA
kprema@umiami.ir.miami.edu

P.H. Bauer
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556 USA
pbauer@mars.ee.nd.edu

ABSTRACT

Delta-operator based implementations can avoid the numerical ill-conditioning usually associated with high speed shift-operator based implementations of discrete-time systems. Moreover, it provides a unified methodology for tackling both continuous- and discrete-time systems. In particular, it has been shown that, delta-operator based balanced realizations can offer superior coefficient sensitivity properties under fixed-point arithmetic. In this work, we address computation of balanced realizations. For this purpose, given a discrete-time system, the relationship between its shift- and delta-operator formulated balanced realizations is presented.

I. INTRODUCTION

Current interest in delta-systems (δ-systems) is due mainly to two reasons: (a) δ-systems provide superior roundoff noise [1-2] and coefficient sensitivity [3-4] properties, and (b) δ-operator makes it possible to treat both continuous-time (CT) and discrete-time (DT) systems in a unified manner [5]. Recent work on δ-operator based implementation of two-dimensional (2-D) DT systems contain the counterpart to the shift-operator (q-operator) based Roesser local state-space (s.s.) model [6]. Balanced (BL) realization of such models and coefficient sensitivity properties were also investigated. Indeed, given a 2-D DT system, under fixed-point (FXP) arithmetic (and mild conditions), Roesser δ-model was shown to be superior to the Roesser q-model. In this paper, we reveal the relationship between BL realizations of Roesser δ- and q-models. This makes it possible to use techniques available for computation of q-BL models for computation of δ-BL models.

II. NOMENCLATURE AND PRELIMINARIES

2.1. Nomenclature

$\Re$, $\Im$, and $\aleph$ denote the reals, complex numbers, and non-negative integers, respectively. $\Re^{q\times p}$ and $\Im^{q\times p}$ are the sets of matrices of size $q \times p$ over $\Re$ and $\Im$, respectively.

I_n is the unit matrix of size $n \times n$; $\mathbf{0}$ is the null matrix of size $q \times p$. A^* and A^T denote the complex conjugate transpose and transpose of matrix $A \in \Re^{q\times p}$; trace$[A]$ and $\lambda_i[A]$ denote its trace and i-th eigenvalue. $\|A\|_F$ is its Fröbenius norm.

In the 1-D case, corresponding q- and δ-systems are related by $\delta = (q-1)/\Delta \iff c = (z-1)/\Delta$. Here, Δ is a positive real constant (usually the sampling time). For 2-D systems, subscripts h and v denote horizontally propagating (h.p.) and vertically propagating (v.p.) subsystems of the corresponding Roesser local s.s. models. n_h and n_v denote the sizes of these h.p. and v.p. subsystems. We use n to denote $n = n_h + n_v$. Δ_h and Δ_v are positive real constants denoting 'sampling times' along h.p. and v.p. directions.

We use ξ to denote $\Delta_h I_{n_h} \oplus \Delta_v I_{n_v} \in \Re^{n\times n}$. Also, I_z and I_c denote $z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$ and $c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n\times n}$, respectively.

Corresponding 2-D q- and δ-systems are related by $\delta_h = (q_h-1)/\Delta_h \iff c_h = (z_h-1)/\Delta_h$ and $\delta_v = (q_v-1)/\Delta_v \iff c_v = (z_v-1)/\Delta_v$. We use subscripts δ and q to differentiate between corresponding δ- and q-systems; for example, s.s. realization of a given DT system is either $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ if implemented based on δ-operator or $\{A_q, B_q, C_q, D_q\}$ if implemented based on q-operator. The following notation is also used: $H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = H(c_h, c_v)\big|_{\substack{c_h=(z_h-1)/\Delta_h \\ c_v=(z_v-1)/\Delta_v}}$ and $G(z_h, z_v)|_{\mathbf{z}\to\mathbf{c}} = G(z_h, z_v)\big|_{\substack{z_h=1+\Delta_h c_h \\ z_v=1+\Delta_v c_v}}$.

Stability studies of q- and δ-systems involve the follow-

ing regions: $\mathcal{U}_q = \{z \in \Im : |z| < 1\}$; $\mathcal{U}_q^2 = \{(z_h, z_v) \in \Im^2 : |z_h| < 1, |z_v| < 1\}$; $\mathcal{U}_\delta = \{c \in \Im : |c + 1/\Delta| < 1/\Delta\}$; $\mathcal{U}_\delta^2 = \{(c_h, c_v) \in \Im^2 : |c_h + 1/\Delta_h| < 1/\Delta_h, |c_v + 1/\Delta_v| < 1\}$. The corresponding distinguished boundaries are denoted with letter $\mathcal{T}$.; $\mathcal{U}. \cup \mathcal{T}.$ is denoted by $\overline{\mathcal{U}}.$. A q-system polynomial with all its roots in $\mathcal{U}_q$ (for the 1-D case) or $\mathcal{U}_q^2$ (for the 2-D case) is said to be *stable*. The corresponding regions for a δ-system polynomial are $\mathcal{U}_\delta$ (for the 1-D case) and $\mathcal{U}_\delta^2$ (for the 2-D case), respectively.

2.2. Preliminaries

First, we provide a brief overview of relevant material.

Roesser q-model. The 2-D dynamical system under consideration is assumed to be linear, shift-invariant, strictly causal, and modeled by a set of first-order vector difference equations over $\Re$. Given such a p-input and q-output system, its n_hh-n_vv Roesser local s.s. model $\{A_q, B_q, C_q, D_q\}$ is of the form [7]

$$\begin{aligned}
\begin{matrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{matrix} &= \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [A_q] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [B_q]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_q]\mathbf{u}(i,j) \\
&\doteq [C_q] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_q]\mathbf{u}(i,j),
\end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. Also, $A_q^{(1)} \in \Re^{n_h \times n_h}$, $A_q^{(4)} \in \Re^{n_v \times n_v}$, etc. Here, $(i,j) \in \aleph^2$ and

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

Usually, $\mathbf{x}^h$ and $\mathbf{x}^v$ are called the *h.p.* and *v.p.* local state vectors of $\{A_q, B_q, C_q, D_q\}$. With no nonessential singularities of the second kind on $\mathcal{T}_q^2$, for BIBO stability, one requires [8]

$$\det[I_z - A_q] \neq 0, \ \forall (z_h, z_v) \in \overline{\mathcal{U}}_q^2. \tag{2.3}$$

Roesser δ-model. To exploit the superior finite wordlength properties of δ-operator implementations, analogous to the 1-D case, in [6], the following operators are defined:

$$\begin{aligned}
\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\Delta_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h}; \\
\delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\Delta_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v},
\end{aligned} \tag{2.4}$$

where Δ_h and Δ_v are two positive real numbers. Hence, the following relationships are applicable:

$$\delta_h = \frac{q_h - 1}{\Delta_h}; \qquad \delta_v = \frac{q_v - 1}{\Delta_v}. \tag{2.5}$$

Using (2.4-5) in (2.1), the following Roesser δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ has been proposed [6]:

$$\begin{aligned}
\begin{matrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{matrix} &= \begin{bmatrix} A_\delta^{(1)} & A_\delta^{(2)} \\ A_\delta^{(3)} & A_\delta^{(4)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + \begin{bmatrix} B_\delta^{(1)} \\ B_\delta^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [A_\delta] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [B_\delta]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= \begin{bmatrix} C_\delta^{(1)} & C_\delta^{(2)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_\delta]\mathbf{u}(i,j) \\
&\doteq [C_\delta] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_\delta]\mathbf{u}(i,j),
\end{aligned} \tag{2.6}$$

where

$$A_\delta = \xi^{-1}(A_q - I_n); \ B_\delta = \xi^{-1}B_q; \ C_\delta = C_q; \ D_\delta = D_q. \tag{2.7}$$

Here, $\xi = [\Delta_h I_{n_h} \oplus \Delta_v I_{n_v}] \in \Re^{n \times n}$. Note that, as opposed to its corresponding Roesser q-model, here, one must also perform the following computations:

$$q_h[\mathbf{x}^h] = \mathbf{x}^h + \Delta_h \cdot \delta_h[\mathbf{x}^h]; \quad q_v[\mathbf{x}^v](i,j) = \mathbf{x}^v + \Delta_v \cdot \delta_v[\mathbf{x}^v]. \tag{2.8}$$

In [6], several properties of this Roesser δ-model (such as, general response equation, transition matrix, characteristic equation, transfer function) are elaborated. Also, it is easy to see that, as for the q-model, 2-D equivalent transformations of the type

$$\begin{bmatrix} \tilde{\mathbf{x}}^h(i,j) \\ \tilde{\mathbf{x}}^v(i,j) \end{bmatrix} = \begin{bmatrix} T^{(1)} & \mathbf{0} \\ \mathbf{0} & T^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} \doteq [T] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}, \tag{2.9}$$

where $T^{(1)} \in \Re^{n_h \times n_h}$ and $T^{(4)} \in \Re^{n_v \times n_v}$ are nonsingular, yield an equivalent 2-D s.s. realization $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$, where

$$\tilde{A}_\delta = TA_\delta T^{-1}; \ \tilde{B}_\delta = TB_\delta; \ \tilde{C}_\delta = C_\delta T^{-1}; \ \tilde{D}_\delta = D_\delta. \tag{2.10}$$

Also, $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ and $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ have the same transfer function. With no nonessential singularities of the second kind on $\mathcal{T}_\delta^2$, for BIBO stability, one requires

$$\det[I_c - A_\delta] \neq 0, \ \forall (c_h, c_v) \in \overline{\mathcal{U}}_\delta^2. \tag{2.11}$$

III. GRAMIANS AND BL REALIZATIONS

3.1. Gramians

For the Roesser q-model, gramians are taken to be natural extensions of the integral expressions of their 1-D counterparts [9-10]. The work in [6] adopts a similar approach in proposing gramians for the δ-operator case as defined in [5]. In what follows, $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ (with gramians P_δ and Q_δ) and $\{A_q, B_q, C_q, D_q\}$ (with gramians P_q and Q_q) denote a given stable 2-D DT system's δ- and q-operator based Roesser models, respectively.

DEFINITION 3.1. [9-10].

1. Gramians of $\{A_q, B_q, C_q, D_q\}$ are

$$\begin{aligned}
P_q &= \frac{1}{(2\pi j)^2} \oint \oint_{\mathcal{T}_q^2} F_q F_q^* \frac{dz_h}{z_h} \frac{dz_v}{z_v}; \\
Q_q &= \frac{1}{(2\pi j)^2} \oint \oint_{\mathcal{T}_q^2} G_q^* G_q \frac{dz_h}{z_h} \frac{dz_v}{z_v},
\end{aligned}$$

where $F_q(z_h, z_v) = (I_z - A_q)^{-1}B_q$ and $G_q(z_h, z_v) = C_q(I_z - A_q)^{-1}$.

2. Gramians of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ are

$$P_\delta = \frac{1}{(2\pi j)^2}\oint_{T_\delta^2} F_\delta F_\delta^* \frac{dc_h}{1+\Delta_h c_h}\frac{dc_v}{1+\Delta_v c_v};$$

$$Q_d d = \frac{1}{(2\pi j)^2}\oint_{T_\delta^2} G_\delta^* G_\delta \frac{dc_h}{1+\Delta_h c_h}\frac{dc_v}{1+\Delta_v c_v},$$

where $F_\delta(c_h, c_v) = (I_c - A_\delta)^{-1}B_\delta$ and $G_\delta(c_h, c_v) = C_\delta(I_c - A_\delta)^{-1}$.

LEMMA 3.1. [6]. The relationship between the above gramians are

$$P_\delta = \frac{1}{\Delta_h\Delta_v}P_q; \qquad Q_\delta = \frac{1}{\Delta_h\Delta_v}\xi Q_q \xi.$$

With appropriate partitions incorporated, this is equivalent to

$$\begin{bmatrix} P_\delta^{(1)} & P_\delta^{(2)} \\ P_\delta^{(3)} & P_\delta^{(4)} \end{bmatrix} = \frac{1}{\Delta_h\Delta_v}\begin{bmatrix} P_q^{(1)} & P_q^{(2)} \\ P_q^{(3)} & P_q^{(4)} \end{bmatrix};$$

$$\begin{bmatrix} Q_\delta^{(1)} & Q_\delta^{(2)} \\ Q_\delta^{(3)} & Q_\delta^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{\Delta_h}{\Delta_v}Q_q^{(1)} & Q_q^{(2)} \\ Q_q^{(3)} & \frac{\Delta_v}{\Delta_h}Q_q^{(4)} \end{bmatrix}.$$

LEMMA 3.2. [6]. The realization $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ obtained with a nonsingular transformation of the type in (2.9-10) yields the gramians $\tilde{P}_\delta = TP_\delta T^*$ and $\tilde{Q}_\delta = T^{-1^*}Q_\delta T^{-1}$. Eigenvalues of $P_\delta Q_\delta$ are invariant under such a transformation. The situation regarding Roesser q-model is completely equivalent.

DEFINITION 3.2. [10]. Roesser δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is said to be *balanced (BL)* if

$$P_\delta^{(1)} = Q_\delta^{(1)} \doteq \Sigma_\delta^{(1)} = \mathrm{diag}\{\sigma_{\delta_1}^{(1)}, \ldots, \sigma_{\delta_{n_h}}^{(1)}\};$$

$$P_\delta^{(4)} = Q_\delta^{(4)} \doteq \Sigma_\delta^{(4)} = \mathrm{diag}\{\sigma_{\delta_1}^{(4)}, \ldots, \sigma_{\delta_{n_v}}^{(4)}\}.$$

We refer to $\sigma_{\delta_i}^{(1)}$, $i = 1, \ldots, n_h$, and $\sigma_{\delta_j}^{(4)}$, $j = 1, \ldots, n_v$, as the *Hankel singular values* of h.p. and v.p. subsystems, repectively. The situation regarding Roesser q-model is completely equivalent.

3.2. Computation of BL Realizations

Computation of gramians and obtaining BL realizations for q-systems have been investigated quite thoroughly. In the 1-D and 2-D separable cases, one may solve Lyapunov equations and use Laub's algorithm [10-11]. In the 2-D nonseparable case, this computation is not that easy; however, several techniques have been developed [10], [12].

In this section, we provide the relationship between BL realizations of corresponding δ- and q-models. This allows all available techniques for gramian computation of q-systems to be utilized for δ-systems as well. To the authors' knowledge, such a relationship is not available even for the 1-D case. Although we concentrate on the 2-D case, a similar argument may be developed for the 1-D case.

For convenience, we use the following notation:
$\{A, B, C, D\} \xrightarrow{T} \{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$: Here, $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$, $\tilde{C} = CT^{-1}$, and $\tilde{D} = D$, where T is of type (2.9-10).
$\{A_q, B_q, C_q, D_q\} \xrightarrow{q\to\delta} \{A_\delta, B_\delta, C_\delta, D_\delta\}$: This is the corresponding δ-system obtained by applying (2.7).
$\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{\delta\to q} \{A_q, B_q, C_q, D_q\}$: This is the corresponding q-system obtained by applying (2.7).

Moreover, we use the following:
$\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$: BL realization of $\{A_q, B_q, C_q, D_q\}$ obtained via $\{A_q, B_q, C_q, D_q\} \xrightarrow{T_q} \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$.
$\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$: BL realization of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ obtained via $\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{T_\delta} \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$.
$\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$: q-system obtained via $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\} \xrightarrow{\delta\to q} \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$.
$\{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}$: δ-system obtained via $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\} \xrightarrow{q\to\delta} \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}$.

LEMMA 3.3. The following relationships are true:

$$\{A_q, B_q, C_q, D_q\} \xrightarrow{T_\delta} \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\};$$

$$\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{T_q} \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}.$$

Proof. Note that, $A_{\delta B2q} = I_n + \xi A_{\delta B} = I_n + \xi T_\delta A_\delta T_\delta^{-1} = I_n + \xi T_\delta \xi^{-1}(A_q - I_n)T_\delta^{-1} = T_\delta A_q T_\delta^{-1}$, since $\xi T_\delta \xi^{-1} = T_\delta$. The remainder may proven in a similar manner. ∎

LEMMA 3.4. The following relationships are true:

$$\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\} \xrightarrow{\xi^{-1/2}} \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\};$$

$$\{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\} \xrightarrow{\xi^{1/2}} \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}.$$

Proof. Note that, $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ has following gramians:

$$P_{\delta B} = \begin{bmatrix} \Sigma_\delta^{(1)} & P_{\delta B}^{(2)} \\ P_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}; \quad Q_{\delta B} = \begin{bmatrix} \Sigma_\delta^{(1)} & Q_{\delta B}^{(2)} \\ Q_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}.$$

Hence, from Lemma 3.1, $\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$ has the following gramians:

$$P_{\delta B2q} = \Delta_h\Delta_v \begin{bmatrix} \Sigma_\delta^{(1)} & P_{\delta B}^{(2)} \\ P_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix};$$

$$Q_{\delta B2q} = \begin{bmatrix} \frac{\Delta_v}{\Delta_h}\Sigma_\delta^{(1)} & Q_{\delta B}^{(2)} \\ Q_{\delta B}^{(3)} & \frac{\Delta_h}{\Delta_v}\Sigma_\delta^{(4)} \end{bmatrix}.$$

To get $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$, we need to simultaneously diagonalize the two pairs $\{\Delta_h\Delta_v\Sigma_\delta^{(1)}, (\Delta_v/\Delta_h)\Sigma_\delta^{(1)}\}$ and $\{\Delta_h\Delta_v\Sigma_\delta^{(4)}, (\Delta_h/\Delta_v)\Sigma_\delta^{(4)}\}$. By applying Laub's algorithm, we get these two transformations to be $\Delta_h^{-1/2}I_{n_h}$ and

$\Delta_v^{-1/2} I_{n_v}$. This proves the first part. The remainder follows in a similar manner. ∎

COROLLARY 3.5. The systems $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$ and $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ are related as follows:

$$A_{\delta B} = \xi^{-1/2}(A_{qB} - I_n)\xi^{-1/2};\ B_{\delta B} = \xi^{-1/2} B_{qB};$$
$$C_{\delta B} = C_{qB}\xi^{-1/2};\ D_{\delta B} = D_{qB}.$$

Proof. Note that, from Lemma 3.4, $A_{\delta B} = \xi^{-1}(A_{\delta B 2q} - I_n) = \xi^{-1}(\xi^{1/2} A_{qB} \xi^{-1/2} - I_n) = \xi^{-1/2}(A_{qB} - I_n)\xi^{-1/2}$. The rest follows in a similar manner. ∎

IV. EXAMPLE

We now consider a stable 3h-3v 2-D separable digital filter.

4.1. Computations

Numerical values are *displayed* via FORMAT SHORT E of MATLAB [13] which was used for all computations. Note that, since system being considered has $A_q^{(3)} = \mathbf{0}$ (instead of $A_q^{(2)} = \mathbf{0}$), relevant equations must be appropriately modified.

Given q-model $\{A_q, B_q, C_q, D_q\}$.

$$A_q^{(1)} = \begin{bmatrix} 0 & 1.0000e+00 & 0 \\ 0 & 0 & 1.0000e+00 \\ 3.8315e-01 & -1.3861e+00 & 1.9067e+00 \end{bmatrix};$$

$$A_q^{(2)} = \begin{bmatrix} -6.8280e-02 & 6.1900e-02 & 6.5400e-03 \\ -2.8100e-02 & 3.9560e-02 & -2.2480e-02 \\ 1.2445e+00 & -5.7092e-01 & 2.0587e+00 \end{bmatrix};$$

$$A_q^{(3)} = \mathbf{0}, \text{ or } \begin{bmatrix} 0 & 1.0000e+00 & 0 \\ A_q^{(4)} & 0 & 0 & 1.0000e+00 \\ 3.8238e-01 & -1.3818e+00 & 1.9025e+00 \end{bmatrix};$$

$$B_q^{(1)} = [0 \quad 0 \quad 1]^T;$$
$$B_q^{(2)} = [0 \quad 0 \quad 1]^T;$$
$$C_q^{(1)} = [1.1410e-02 \quad -5.4000e-03 \quad 1.9560e-02];$$
$$C_q^{(2)} = [1.1640e-02 \quad -5.4500e-03 \quad 1.9600e-02];$$
$$D_q = [9.4300e-03].$$

BL q-model $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$.

$$A_{qB}^{(1)} = \begin{bmatrix} 8.6478e-01 & 2.6806e-01 & -3.4799e-02 \\ -2.6806e-01 & 5.8766e-01 & 3.8402e-01 \\ -3.4797e-02 & -3.8401e-01 & 4.5427e-01 \end{bmatrix};$$

$$A_{qB}^{(2)} = \begin{bmatrix} 4.2940e-01 & -3.3765e-01 & 1.2689e-01 \\ 3.3771e-01 & -2.6511e-01 & 1.0134e-01 \\ 1.2732e-01 & -9.7518e-02 & 3.2423e-02 \end{bmatrix};$$

$$A_{qB}^{(3)} = \mathbf{0};$$

$$A_{qB}^{(4)} = \begin{bmatrix} 8.6486e-01 & 2.6760e-01 & -3.4949e-02 \\ -2.6760e-01 & 5.8692e-01 & 3.8661e-01 \\ -3.4952e-02 & -3.8661e-01 & 4.5071e-01 \end{bmatrix};$$

$$B_{qB}^{(1)} = [6.3568e-02 \quad 4.9879e-02 \quad 1.8565e-02]^T;$$
$$B_{qB}^{(2)} = [6.5595e-01 \quad 5.1555e-01 \quad 1.9416e-01];$$
$$C_{qB}^{(1)} = [6.5590e-01 \quad -5.1574e-01 \quad 1.9341e-01];$$
$$C_{qB}^{(2)} = [6.3592e-02 \quad -4.9875e-02 \quad 1.8540e-02];$$
$$D_{qB} = [9.4300e-03].$$

Corresponding δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$. Let us select $\Delta_h = Delta_v = 2.5000e-01$. Acoordingly, we get

$$A_\delta^{(1)} = \begin{bmatrix} -4.0000e+00 & 4.0000e+00 & 0 \\ 0 & -4.0000e+00 & 4.0000e+00 \\ 1.5326e+00 & -5.5444e+00 & 3.6268e+00 \end{bmatrix};$$

$$A_\delta^{(2)} = \begin{bmatrix} -2.7312e-01 & 2.4760e-01 & 2.6160e-02 \\ -1.1240e-01 & 1.5824e-01 & -8.9920e-02 \\ 4.9780e+00 & -2.2837e+00 & 8.2348e+00 \end{bmatrix};$$

$$A_\delta^{(3)} = \mathbf{0};$$

$$A_\delta^{(4)} = \begin{bmatrix} -4.0000e+00 & 4.0000e+00 & 0 \\ 0 & -4.0000e+00 & 4.0000e+00 \\ 1.5295e+00 & -5.5272e+00 & 3.6100e+00 \end{bmatrix};$$

$$B_\delta^{(1)} = [0 \quad 0 \quad 4]^T;$$
$$B_\delta^{(2)} = [0 \quad 0 \quad 4]^T;$$
$$C_\delta^{(1)} = [1.1410e-02 \quad -5.4000e-03 \quad 1.9560e-02];$$
$$C_\delta^{(2)} = [1.1640e-02 \quad -5.4500e-03 \quad 1.9600e-02];$$
$$D_\delta = [9.4300e-03].$$

BL δ-model $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$.

$$A_{\delta B}^{(1)} = \begin{bmatrix} -5.4089e-01 & 1.0722e+00 & -1.3919e-01 \\ -1.0722e+00 & -1.6494e+00 & 1.5361e+00 \\ -1.3919e-01 & -1.5361e+00 & -2.1829e+00 \end{bmatrix};$$

$$A_{\delta B}^{(2)} = \begin{bmatrix} 1.7176e+00 & -1.3506e+00 & 5.0755e-01 \\ 1.3508e+00 & -1.0604e+00 & 4.0537e-01 \\ 5.0926e-01 & -3.9007e-01 & 1.2969e-01 \end{bmatrix};$$

$$A_{\delta B}^{(3)} = \mathbf{0};$$

$$A_{\delta B}^{(4)} = \begin{bmatrix} -5.4054e-01 & 1.0704e+00 & -1.3980e-01 \\ -1.0704e+00 & -1.6523e+00 & 1.5464e+00 \\ -1.3981e-01 & -1.5464e+00 & -2.1971e+00 \end{bmatrix};$$

$$B_{\delta B}^{(1)} = [1.2714e-01 \quad 9.9759e-02 \quad 3.7129e-02]^T;$$
$$B_{\delta B}^{(2)} = [1.3119e+00 \quad 1.0311e+00 \quad 3.8833e-01]^T;$$
$$C_{\delta B}^{(1)} = [1.3118e+00 \quad -1.0315e+00 \quad 3.8682e-01];$$
$$C_{\delta B}^{(2)} = [1.2718e-01 \quad -9.9750e-02 \quad 3.7080e-02];$$
$$D_{\delta B} = [9.4300e-03].$$

4.2. Simulations

Normalized frequency response of $\{A_q, B_q, C_q, D_q\}$ is $H_q\left(e^{j\omega_1}, e^{j\omega_2}\right)$ and that of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is $H_\delta((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v)$. Frequency responses are evaluated on $\mathcal{G}^2 \doteq \{(\omega_1, \omega_2) \in \Re^2 : \omega_i = n_i \times \pi/N,\ n_i = [-N : 1 : N],\ i = 1, 2\}$ with $N = 32$. For comparison purposes, the following measure was also evaluated: For $(z_1, z_2) = (e^{j\omega_1}, e^{j\omega_2})$ and $(c_1, c_2) = ((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v)$,

$$E_{\max} \doteq \begin{cases} \max_{\mathcal{G}^2} \left| H(z_1, z_2) - \hat{H}(z_1, z_2) \right|, & \text{for } q\text{-models;} \\ \max_{\mathcal{G}^2} \left| H(c_1, c_2) - \hat{H}(c_1, c_2) \right|, & \text{for } \delta\text{-models.} \end{cases}$$

Here, H denotes the 'ideal' frequency response where each 'coefficient is represented in 'infinite' precision; $\hat{H}$ denotes the 'actual' frequency response where each coefficient is represented in finite precision.

Fig. (1) shows $E_{\max}$ versus number of fractional bits where each coefficient is represented in FXP and its fractional part is truncated at different lengths; integral part is represented exactly. Advantage gained by BL δ-model over BL q-model is about 3 bits.

Fig. (2) shows $E_{\max}$ versus total number of bits where each coefficient is represented in FXP and its total (integral+fractional) number of bits is truncated at different lengths. Advantage gained by BL δ-model over BL q-model is only about 1 bit. This modest improvement is due to the large Δ_h and Δ_v being used. More dramatic improvements require smaller Δ_h and Δ_v [6]. But, this makes δ-model's coefficients to occupy a larger dynamic range. To circumvent this, we believe, careful scaling of filter coefficients is necessary. We are currently investigating this possibility.

Fig. (3) shows $E_{\max}$ versus number of mantissa bits where each coefficient is represented in FLP and its number of mantissa bits is truncated at different lengths. Of course, in FLP, dynamic range is usually of no threat.

V. CONCLUSION

In this work, we have presented the relationship between BL realizations of corresponding δ- and q-models. This, in turn, addresses computation of gramians and BL realizations of δ-models.

In the FXP case, δ-model is better whenever $\Delta_h < 1$ and $\Delta_v < 1$ [6]. However, this choice must be carefully done since, in FXP, δ-models tend to occupy a larger dynamic range. The authors are currently investigating the possibility of incorporating scaling of coefficients so that low values of Δ_h and Δ_v may be used to expose and exploit the advantages of δ-systems. In the FLP case, such a limitation does not usually arise, and δ-models are better whenever the system matrix eigenvalues lie within a certain region called the *MG-region* [14]. This condition is typically true for high Q, narrowband digital filters operating under high sampling rates. These observations indicate that, in FLP, for comparative performance (with respect to coefficient sensitivity), δ-models require a shorter mantissa length. The ensuing implications regarding low power consumption, low cost and weight, and high speed cannot be overemphasized. The authors are currently completing work regarding quantization noise properties of the δ-model developed, where, as in 1-D case, improvements over the corresponding q-model are expected.

We must mention that certain difficulties regarding limit cycles are inherent in δ-systems when FXP arithmetic is used [15]. However, this problem is, for all practical purposes, nonexistent in FLP arithmetic. Hence, in our opinion, for FLP high performance applications, the δ-model developed provides an extremely attractive solution that avoids numerical ill-conditioning typically associated with high speed q-systems.

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proc. IEEE,* vol. 80, pp. 240-259, 1992.

[2] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.,* vol. 41, pp. 629-637, Feb. 1993.

[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proc. 1990 IEEE Conf. Decision and Cont. (CDC'90),* pp. 954-959, Honolulu, HI, Dec. 1990.

[4] K. Premaratne, R. Salvi, N.R. Habib, and J.P. Le Gall, "Delta-operator formulated discrete-time equivalents of continuous-time systems," *IEEE Trans. Auto. Cont.,* vol. 39, pp. 581-585, Mar. 1994.

[5] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1990.

[6] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer, "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties," *Proc. 37th Midwest Symp. Circ. Syst. (MWSCS'94),* Lafayette, LA, Aug. 1994.

[7] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Trans. Auto. Cont.,* vol. AC-20, pp. 1-10, Feb. 1975.

[8] E.I. Jury, "Stability of multidimensional systems and other related problems," Chapter 3 in *Multidimensional Systems, Techniques, and Applications,* New York, NY: Marcel Dekkar, 1986.

[9] W.-S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proc. 1986 IEEE Int. Symp. Circ. Syst. (ISCAS'86),* vol. 1, pp. 79-82, 1986.

[10] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Trans. Circ. Syst.,* vol. CAS-37, pp. 1116-1132, Sept. 1990.

[11] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Auto. Cont.,* vol. AC-32, pp. 115-122, Feb. 1987.

[12] W.-S. Lu, H.-P. Wang, and A. Antoniou, "An efficient method for the evaluation of the controllability and observability gramians of 2-D digital filters and systems," *IEEE Trans. Circ. Syst.—II. Anal. Dig. Sig. Proc.,* vol. 39, pp. 695-704, Oct. 1992.

[13] *MATLAB,* ver. 4.2a, Natick, MA: The MathWorks Inc.

[14] K. Premaratne, M.M. Ekanayake, J. Suarez, and P.H. Bauer, "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties" (detailed version of [6]), *IEEE Trans. Sig. Proc.,* in review, 1995.

[15] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proc. 1994 IEEE Int. Symp. Circ. Syst. (ISCAS'94),* London, UK, vol. 2, pp. 461-464, May 1994.

ACKNOWLEDGEMENT

The work of K.P. and P.H.B. were partially supported by the US Office of Naval Research (ONR) through grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

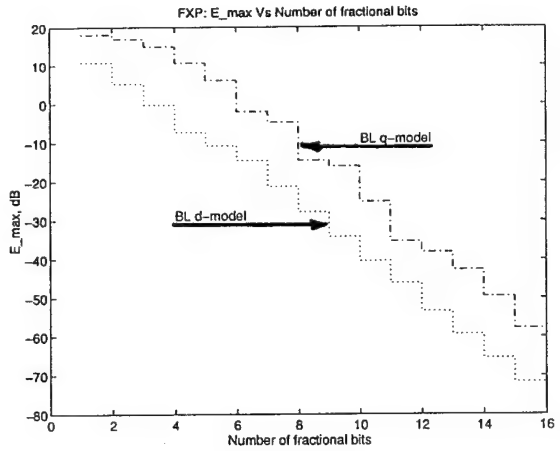


Figure (1)

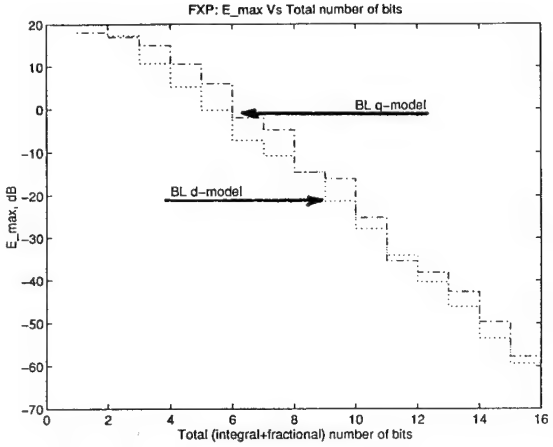


Figure (2)

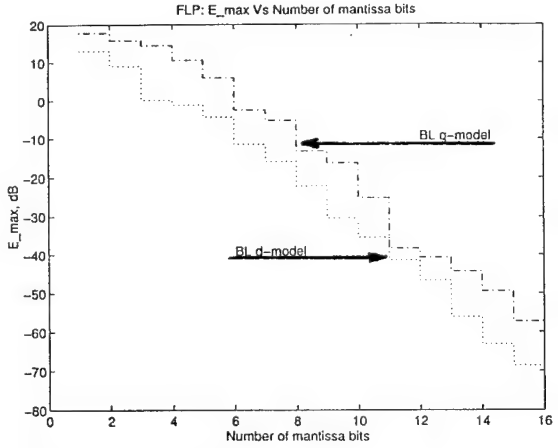


Figure (3)

Proceedings of the 37th Midwest Symposium on Circuits and Systems (MWSCS'94), Lafayette, LA, August 1994.

Two-Dimensional Delta-Operator Formulated Discrete-Time Systems: State-Space Realization and Its Coefficient Sensitivity Properties

K. Premaratne, J. Suarez, and M.M. Ekanayake
Department of E&CE
University of Miami
Coral Gables, FL 33124 USA

P.H. Bauer
Department of EE
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556 USA

Abstract—By developing the δ-operator analog of the Roesser model, state-space realization of two- and multi-dimensional δ-systems is investigated. The corresponding notions of gramians and balanced realization are also defined. It is shown that, discrete-time system implementation using this model can yield superior coefficient sensitivity properties.

I. Introduction

Judging by its performance in the one-dimensional (1-D) case [2], [5-6], one is led to expect superior coefficient sensitivity and roundoff noise performance with δ-operator implementation of two-dimensional (2-D) and multi-dimensional (m-D) discrete-time (DT) systems. With this in mind, δ-operator analog of the q-operator Roesser local state-space (s.s.) model [12] is developed. We also propose the notions of gramians and balanced (BL) realization. As expected, realization using this model can provide superior coefficient sensitivity properties.

II. Nomenclature and Preliminaries

A. Nomenclature

$\Re$: Reals; $\Im$: Complex numbers; $\Re^{q\times p}$, $\Im^{q\times p}$: Matrices of size $q \times p$ over $\Re$ and $\Im$; I_n: $n \times n$ unit matrix; A^*, trace$[A]$, $\|A\|_F$: Conjugate transpose, trace, and Fröbenius norm of matrix A; $\mathbf{e}_i^{(n)}$: Unit vector in $\Re^n$ with 1 on the i-th row; $E_{i,j}^{q\times p} = \mathbf{e}_i^{(q)}\mathbf{e}_j^{(p)^*} \in \Re^{q\times p}$; $\overline{U}_{q\times p} = \sum_{i=1}^{q}\sum_{j=1}^{p} E_{i,j}^{(q\times p)} \otimes E_{i,j}^{(q\times p)} \in \Re^{q^2\times p^2}$.

For q- and δ-systems, we use the indeterminates z and c, respectively. For 1-D systems, $\delta = (q-1)/\tau \Longleftrightarrow c = (z-1)/\tau$, where τ is a positive real constant, usually the sampling time. Let $\overline{\mathcal{U}}_\delta^2 = \{(c_h, c_v) \in \Im^2 : |c_h + 1/\tau_h| \le 1/\tau_h, |c_v + 1/\tau_v| \le 1\}$. $\mathcal{T}_\delta^2$ is its boundary. The corresponding q-system regions are denoted with the subscript q.

K.P. and P.H.B. gratefully acknowledge the support received from the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

B. Preliminaries

Consider a linear, shift-invariant, strictly causal, p-input q-output 2-D DT system. Its n_hh-n_vv Roesser local s.s. model $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ takes the form [12]:

$$\begin{aligned}\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j)\end{bmatrix} &= [\hat{A}]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [\hat{B}]\mathbf{u}(i,j);\\ \mathbf{y}(i,j) &\doteq [\hat{C}]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [\hat{D}]\mathbf{u}(i,j),\end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. $\mathbf{x}^h$ and $\mathbf{x}^v$ are the h.p. and v.p. local state vectors. Take $n = n_h + n_v$. Also,

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j); \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

In what follows, we use matrix partitioning that conform to $A \doteq \begin{bmatrix}\hat{A}^{(1)} & \hat{A}^{(2)}\\ \hat{A}^{(3)} & \hat{A}^{(4)}\end{bmatrix}$, $B \doteq \begin{bmatrix}\hat{B}^{(1)}\\ \hat{B}^{(2)}\end{bmatrix}$, and $C \doteq \left[\hat{C}^{(1)}, \hat{C}^{(2)}\right]$. The corresponding 2-D characteristic equation and transfer function are

$$\begin{aligned}\det[I_z - \hat{A}] &= \det[z_h I_{n_h} \oplus z_v I_{n_v} - \hat{A}];\\ \hat{H}(z_h, z_v) &= \hat{C}(I_z - \hat{A})^{-1}\hat{B} + \hat{D},\end{aligned} \tag{2.3}$$

where $z_h, z_v \in \Im$ $I_z \doteq z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$. With no nonessential singularities of the second kind (NSSK) on $\mathcal{T}_q^2$, $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ is BIBO stable iff [3]

$$\det[I_z - \hat{A}] \neq 0, \ \forall (z_h, z_v) \in \overline{\mathcal{U}}_q^2. \tag{2.4}$$

III. 2-D δ-Model

A. Local s.s. model

Analogous to the 1-D case, define $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as

$$\begin{aligned}\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\tau_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\tau_h};\\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\tau_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\tau_v}.\end{aligned} \tag{3.1}$$

Here τ_h and τ_v are positive real constants denoting the 'sampling times' along h.p. and v.p. directions, respectively. Note that

$$q_h = 1 + \tau_h \delta_h; \quad q_v = 1 + \tau_v \delta_v, \tag{3.2}$$

and letting $\tau = [\tau_h I_{n_h} \oplus \tau_v I_{n_v}] \in \Re^{n \times n}$,

$$\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} = I_n + \tau \begin{bmatrix} \delta_h I_{n_h} & 0 \\ 0 & \delta_v I_{n_v} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}. \tag{3.3}$$

Using (3.3) in (2.1), we get

$$\begin{aligned} \begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} &\doteq [A] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &\doteq [C] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j), \end{aligned} \tag{3.4}$$

where $A \doteq \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix}$, $B \doteq \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix}$, and $C \doteq \begin{bmatrix} C^{(1)}, C^{(2)} \end{bmatrix}$. In addition, we need to perform

$$q_h[\mathbf{x}^h] = \mathbf{x}^h + \tau_h \cdot \delta_h[\mathbf{x}^h]; \; q_v[\mathbf{x}^v] = \mathbf{x}^v + \tau_v \cdot \delta_v[\mathbf{x}^v]. \tag{3.5}$$

Here,

$$\hat{A} = I_n + \tau A; \; \hat{B} = \tau B; \; \hat{C} = C; \; \hat{D} = D. \tag{3.6}$$

B. Properties of the 2-D δ-model

Most of the following properties may be derived in a manner that is exactly analogous to that in [12].

The *transition matrix* $A^{i,j}$ of the δ-model, may be recursively computed from

$$A^{i,j} = \begin{cases} \mathbf{0}, (i,j) = (0,0); \\ [I_{n_h} \oplus I_{n_v}], (i,j) = (0,0); \\ \begin{bmatrix} I_{n_h} & 0 \\ 0 & 0 \end{bmatrix} + \tau \begin{bmatrix} A^{(1)} & A^{(2)} \\ 0 & 0 \end{bmatrix}, (i,j) = (1,0); \\ \begin{bmatrix} 0 & 0 \\ 0 & I_{n_v} \end{bmatrix} + \tau \begin{bmatrix} 0 & 0 \\ A^{(3)} & A^{(4)} \end{bmatrix}, (i,j) = (0,1); \\ A^{1,0} A^{i-1,j} + A^{0,1} A^{i,j-1}, \text{ elsewhere.} \end{cases} \tag{3.7}$$

The general response of the δ-model is

$$\begin{aligned} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} &= \sum_{k=0}^{j} A^{i,j-k} \begin{bmatrix} \mathbf{x}^h(0,k) \\ 0 \end{bmatrix} \\ &\quad + \sum_{h=0}^{i} A^{i-j,j} \begin{bmatrix} 0 \\ \mathbf{x}^v(h,0) \end{bmatrix} + \mathbf{f}(\mathbf{u}), \end{aligned} \tag{3.8}$$

where $\mathbf{f}(\mathbf{u}) = \sum_{(0,0) \le (h,k) < (i,j)} (A^{i-h-1,j-k} \tau \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i-h,j-k-1} \tau \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix}) \mathbf{u}(h,k)$.

Let $I_c \doteq c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n \times n}$. Then, the 2-D δ-model's characteristic equation and transfer function are

$$\begin{aligned} \det[I_c - A] &= \frac{1}{\det[\tau]} \det[I_z - \hat{A}]|_{\mathbf{z} \to \mathbf{c}}; \\ H(c_h, c_v) &= \hat{H}(z_h, z_v)|_{\mathbf{z} \to \mathbf{c}}, \end{aligned} \tag{3.9}$$

where

$$z_h = 1 + \tau_h c_h; \quad z_v = 1 + \tau_v c_v. \tag{3.10}$$

From now on, the variable transformation in (3.10) is denoted by $\mathbf{c} \to \mathbf{z}$ or $\mathbf{z} \to \mathbf{c}$ whatever is appropriate.

Nonsingular transformations of the type

$$\begin{bmatrix} \tilde{\mathbf{x}}^h(i,j) \\ \tilde{\mathbf{x}}^v(i,j) \end{bmatrix} = [T] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}, \tag{3.11}$$

where $T \doteq [T^{(1)} \oplus T^{(4)}]$, yield the equivalent 2-D s.s. realization $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$. Here,

$$\tilde{A} = TAT^{-1}; \; \tilde{B} = TB; \; \tilde{C} = CT^{-1}; \; \tilde{D} = D. \tag{3.12}$$

With no NSSK on $\mathcal{T}_\delta^2$, $\{A, B, C, D\}$ is BIBO stable iff

$$\det[I_c - A] \neq 0, \; \forall (c_h, c_v) \in \overline{\mathcal{U}}_\delta^2. \tag{3.13}$$

C. Gramians

The gramians of 2-D q-systems are taken to be natural extensions of the integral expressions of their 1-D counterparts [11]. We will adopt a similar approach. In what follows, we consider the 1-D (or 2-D) stable δ-system $\{A, B, C, D\}$ with gramians P and Q. The corresponding q-system is $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ with gramians $\hat{P}$ and $\hat{Q}$.

1-D case. The gramians are defined in [10].

Definition 3.1. [10]. The gramians are the solutions to the Lyapunov equations

$$\begin{aligned} AP + PA^* + \tau \cdot APA^* &= -BB^*; \\ A^*Q + QA + \tau \cdot A^*QA &= -C^*C. \end{aligned}$$

Lemma 3.1. The gramians satisfy the integral expressions

$$P = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} FF^* \frac{dc}{1 + \tau c}; \; Q = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} G^*G \frac{dc}{1 + \tau c},$$

where $F(c) \doteq (cI_n - A)^{-1} B$ and $G(c) \doteq C(cI_n - A)^{-1}$. Moreover, $\hat{P} = \tau P$ and $\hat{Q} = Q/\tau$.

Proof. Substitute $\hat{A} = I_n + \tau A$, $\hat{B} = \tau B$, $\hat{C} = C$, and $\hat{D} = D$ [10] in the equations in Definition 3.1, and note the integral expressions for P and Q in [8]. ■

2-D case. With Lemma 3.1 in mind, we have

Definition 3.2. The gramians are

$$P = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} FF^* \frac{dc_h}{1+\tau_h c_h} \frac{dc_v}{1+\tau_v c_v};$$

$$Q = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} G^*G \frac{dc_h}{1+\tau_h c_h} \frac{dc_v}{1+\tau_v c_v},$$

where $P \doteq \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix}$ and $Q \doteq \begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix}$. Also, $F(c_h, c_v) \doteq (I_c - A)^{-1}B = [\mathbf{f}_1, \ldots, \mathbf{f}_n]^*$ and $G(c_h, c_v) \doteq C(I_c - A)^{-1} = [\mathbf{g}_1, \ldots, \mathbf{g}_n]$.

Remarks.

1. Note that, $(I_c - A)^{-1}|_{\mathbf{c}\to\mathbf{z}} = (I_z - \hat{A})^{-1}\tau$, and

$$F|_{\mathbf{c}\to\mathbf{z}} = \hat{F};\ G|_{\mathbf{c}\to\mathbf{z}} = \hat{G}\cdot\tau. \tag{3.14}$$

2. Definition 3.2 is completely analogous to the 1-D and 2-D q-systems [7], [11].

Lemma 3.2. $\hat{P} = \tau_h\tau_v P$ and $\hat{Q} = \tau_h\tau_v\tau^{-1}Q\tau^{-1}$.

Proof. Consider P in Definition 3.2. Use $\mathbf{c} \to \mathbf{z}$, (3.14), and definition of gramians for 2-D q-systems [11]. ■

The following are in complete analogy with 2-D q-systems.

Lemma 3.3. The gramians may be represented as

$$P = \frac{1}{\tau_h\tau_v} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} M_{i,j}M_{ij}^*;$$

$$Q = \frac{1}{\tau_h\tau_v}\tau \cdot \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} A^{i,j^*}C^*CA^{i,j} \cdot \tau,$$

where, for $(i,j) = (0,0)$, $M_{i,j} = \mathbf{0}$, and, for $(i,j) > (0,0)$, $M_{ij} = A^{i-1,j}\tau \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i,j-1}\tau \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix}$.

Lemma 3.4. Let $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ with gramians $\tilde{P}$ and $\tilde{Q}$ be an equivalent system as in (3.10-11). Then, $\tilde{P} = TPT^*$ and $\tilde{Q} = T^{-1^*}QT^{-1}$. Moreover, the eigenvalues of PQ and $\tilde{P}\tilde{Q}$ are invariant.

Definition 3.3. $\{A, B, C, D\}$ is said to be *balanced* if $P^{(1)} = Q^{(1)} \doteq \Sigma^{(1)} = \mathrm{diag}\{\sigma_1^{(1)}, \sigma_2^{(1)}, \ldots, \sigma_{n_h}^{(1)}\}$ and $P^{(4)} = Q^{(4)} \doteq \Sigma^{(4)} = \mathrm{diag}\{\sigma_1^{(4)}, \sigma_2^{(4)}, \ldots, \sigma_{n_v}^{(4)}\}$.

If the diagonal submatrices of P and Q are each positive definite (p.d.), a BL realization may be obtained [4]. Regarding this, we have

Lemma 3.5. Local reachability and observability of $\{A, B, C, D\}$ and $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ are equivalent. Moreover, when $\{A, B, C, D\}$ is locally reachable and observable, $P^{(1)}$, $P^{(4)}$, $Q^{(1)}$, and $Q^{(4)}$ are each p.d.

Separable systems. A separable (in denominator) 2-D q-system will have $\hat{A}^{(2)} = \mathbf{0}$ (and/or $\hat{A}^{(3)} = \mathbf{0}$) and all off-diagonal submatrices of $\hat{P}$ and $\hat{Q}$ are zero. The diagonal submatrices may be computed through two pairs of Lyapunov equations [11]. Clearly, a separable 2-D q-system yields a separable 2-D δ-system.

Theorem 3.6. Let $\{A, B, C, D\}$ be separable with $A^{(2)} = 0$. Then, $P^{(2)} = Q^{(2)} = \mathbf{0}$ and $P^{(3)} = Q^{(3)} = 0$, and

$$\begin{aligned}
&A^{(1)}P^{(1)} + P^{(1)}A^{(1)^*} + \tau_h A^{(1)}P^{(1)}A^{(1)^*} \\
&\quad = -B^{(1)}B^{(1)^*}/\tau_v; \\
&A^{(1)^*}Q^{(1)} + Q^{(1)}A^{(1)} + \tau_h A^{(1)^*}Q^{(1)}A^{(1)} \\
&\quad = -\begin{bmatrix} C^{(1)} & R^{(4)}A^{(3)} \end{bmatrix}^* \begin{bmatrix} C^{(1)} & R^{(4)}A^{(3)} \end{bmatrix}/\tau_v; \\
&A^{(4)}P^{(4)} + P^{(4)}A^{(4)^*} + \tau_v A^{(4)}P^{(4)}A^{(4)^*} \\
&\quad = -\begin{bmatrix} B^{(2)} & A^{(3)}S^{(1)} \end{bmatrix} \begin{bmatrix} B^{(2)} & A^{(3)}S^{(1)} \end{bmatrix}^*/\tau_h; \\
&A^{(4)^*}Q^{(4)} + Q^{(4)}A^{(4)} + \tau_v A^{(4)^*}Q^{(4)}A^{(4)} \\
&\quad = -C^{(2)^*}C^{(2)}/\tau_h.
\end{aligned}$$

Here, $R^{(4)^*}R^{(4)} \doteq \tau_h\tau_v Q^{(4)}$ and $S^{(1)}S^{(1)^*} \doteq \tau_h\tau_v P^{(1)}$.

IV. Coefficient Sensitivity

By generalizing a certain sensitivity measure, Lutz and Hakimi [9] have addressed sensitivity minimization of MIMO 1-D CT systems. The SISO 2-D q-operator case is in [7]. In what follows, we study the coefficient sensitivity of the 2-D δ-model in section III. We follow a more direct approach using Kronecker product formulation and, hence, the results are applicable to the more general MIMO case. Using [1], we may show

$$S_A(c_h, c_v) = [I_n \otimes G]\cdot\overline{U}_{n\times n}\cdot[I_n \otimes F] \tag{4.1}$$

$$S_B(c_h, c_v) = [I_n \otimes G]\cdot\overline{U}_{n\times p} \tag{4.2}$$

$$S_C(c_h, c_v) = \overline{U}_{q\times n}\cdot[I_n \otimes F] \tag{4.3}$$

$$S_D(c_h, c_v) = \overline{U}_{q\times p} \tag{4.4}$$

Lemma 4.1. The quantities in (4.1-4.4) are given as

$$S_A = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} [\mathbf{f}_1^* \quad \cdots \quad \mathbf{f}_n^*];$$

$$S_B = \begin{bmatrix} \mathbf{g}_1^{(1)} & \cdots & \mathbf{g}_1^{(p)} \\ \vdots & \ddots & \vdots \\ \mathbf{g}_n^{(1)} & \cdots & \mathbf{g}_n^{(p)} \end{bmatrix};$$

$$S_C = \begin{bmatrix} \mathbf{f}_1^{(1)^*} & \cdots & \mathbf{f}_n^{(1)^*} \\ \vdots & \ddots & \vdots \\ \mathbf{f}_1^{(q)^*} & \cdots & \mathbf{f}_n^{(q)^*} \end{bmatrix};$$

$$S_D = \begin{bmatrix} E_{1,1} & \cdots & E_{1,p} \\ \vdots & \ddots & \vdots \\ E_{q,1} & \cdots & E_{q,p} \end{bmatrix}.$$

Here, $\mathbf{f}_i^{(j)^*}$ denotes a $(q \times p)$ null matrix except its j-th row which is $\mathbf{f}_i^*$ and $\mathbf{g}_i^{(j)}$ denotes a $(q \times p)$ null matrix except its j-th column which is $\mathbf{g}_i$.

Proof. This may be shown through the results in [1] and simple yet tedious algebraic manipulations. ■

Corollary 4.2. The quantities S_A, S_B, S_C, and S_D of the δ-model and the quantities $\hat{S}_{\hat{A}}$, $\hat{S}_{\hat{B}}$, $\hat{S}_{\hat{C}}$, and $\hat{S}_{\hat{D}}$ of the corresponding q-model are related by $S_A|_{\mathbf{c}\to\mathbf{z}} = \tau\hat{S}_{\hat{A}}$, $S_B|_{\mathbf{c}\to\mathbf{z}} = \tau\hat{S}_{\hat{B}}$, $S_C|_{\mathbf{c}\to\mathbf{z}} = \hat{S}_{\hat{C}}$, and $S_D|_{\mathbf{c}\to\mathbf{z}} = \hat{S}_{\hat{D}}$, where $\tau = \tau_h I_{n_h q} \oplus \tau_v I_{n_v q} \in \Re^{nq\times nq}$.

Proof. Apply (3.14) to Lemma 4.1. ■

To proceed further, we utilize the following

Definition 4.1. Let $H(c_h, c_v)$ be a bivariate matrix-valued function that is analytic on $\mathcal{T}_\delta^2$. Then,

$$\|H(c_h, c_v)\|_p^p \doteq \frac{1}{(2\pi\jmath)^2} \oint_{\mathcal{T}_q^2} \|H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\|_F^p \frac{dz_h}{z_h}\frac{dz_v}{z_v}.$$

Remark. This norm is extensively utilized in related work [7] due mainly to the fact that it leads to tractable results. This, and our desire to make a comparison with the corresponding q-model, are the primary reasons for its use here.

We now define the absolute sensitivity measure

$$M \doteq \|S_A\|_1^2 + \frac{1}{p}\|S_B\|_2^2 + \frac{1}{q}\|S_C\|_2^2 + \frac{1}{pq}\|S_D\|_2^2. \quad (4.5)$$

Remarks.

1. The use of different norms is for mathematical feasibility and tractability [7], [5].
2. The weights associated with each term in (4.5) may be thought of as *averaging factors per input/output.*
3. Due to (3.5), M should contain $\|S_{\tau_h}\|$ and $\|S_{\tau_v}\|$. However, we assume that τ_h and τ_v are selected such that each possess exact binary representations. Hence, these additional terms are neglected.

Using an argument similar to that in [7], one may show the following:

$$\|S_A\|_1^2 \le \text{trace}[\hat{P}] \cdot \text{trace}[\tau\hat{Q}\tau] \quad (4.6)$$

$$\|S_B\|_2^2 = p \cdot \text{trace}[\tau\hat{Q}\tau] \quad (4.7)$$

$$\|S_c\|_2^2 = q \cdot \text{trace}[\hat{P}] \quad (4.8)$$

$$\|S_D\|_2^2 = pq \quad (4.9)$$

Combining (4.5) with (4.6-9), we get

$$M \le \overline{M} \doteq (\text{trace}[\hat{P}] + 1)(\text{trace}[\tau\hat{Q}\tau] + 1). \quad (4.10)$$

It is customary to perform a minimization of $\overline{M}$. Hence, one attempts to characterize those $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ that are 'bound optimal' with respect to M. Analogous to 2-D q-systems case [7], one may for instance show that a BL realization (modulo an orthogonal nonsingular transformation) is 'bound optimal' with respect to M.

Compared to a q-system, its δ-system counterpart yields a smaller $\overline{M}$ whenever $\text{trace}[\hat{Q}] > \text{trace}[\tau\hat{Q}\tau]$, that is,

$$(1 - \tau_h^2) \cdot \text{trace}[\hat{Q}^{(1)}] + (1 - \tau_v^2) \cdot \text{trace}[\hat{Q}^{(4)}] > 0. \quad (4.11)$$

Note that, with the local reachability and observability assumption of $\{A, B, C, D\}$, p.d. of $Q^{(1)}$ and $Q^{(4)}$ (and hence of $\hat{Q}^{(1)}$ and $\hat{Q}^{(4)}$) are guaranteed. Thus, (4.11) is satisfied if $\tau_h < 1$ and $\tau_v < 1$.

VII. Conclusion

We have developed the δ-operator analog of the Roesser local s.s. model. Notions of gramians and BL realization are also proposed. As is expected, under mild conditions, this model offers superior coefficient sensitivity properties.

References

[1] J.W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circ. Syst.*, vol. CAS-25, pp. 772-781, Sept. 1978.

[2] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proc. IEEE,* vol. 80, pp. 240-259, 1992.

[3] E.I. Jury, "Stability of multidimensional systems and other related problems," in *Multidimensional Systems, Techniques, and Applications,* S.G. Tzafestas, Ed., New York: Marcel Dekkar, 1986.

[4] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Auto. Cont.*, vol. AC-32, pp. 115-122, Feb. 1987.

[5] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proc. CDC'90,* Honolulu, Dec. 1990, pp. 954-959.

[6] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.*, vol. 41, pp. 629-637, Feb. 1993.

[7] T. Lin, M. Kawamata, and T. Higuchi, "Minimization of sensitivity of 2-D systems and its relation to 2-D balanced realizations," *Proc. ISCAS'87,* Philadelphia, May 1987, vol. 2, pp. 710-713.

[8] W.S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proc. ISCAS'86,* 1986, vol. 1, pp. 79-82.

[9] W.J. Lutz and S.L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circ. Syst.*, vol. 35, pp. 1114-1122, Sept. 1988.

[10] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs: Prentice-Hall, 1990.

[11] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Trans. Circ. Syst.*, vol. CAS-37, pp. 1116-1132, Sept. 1990.

[12] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Trans. Auto. Cont.*, vol. AC-20, pp. 1-10, Feb. 1975.

VII. Appendix B: Papers Partly Related to Grant

Shift-Variant m-D Systems and Singularities on T^m: Implications for Robust Stability

S. A. Yost
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana
(219) 631-7850
yost.2@nd.edu

P. H. Bauer
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana
(219) 631-8015
pbauer@mars.ee.nd.edu

Submitted as a Transactions Brief to **IEEE Transactions on Circuits and Systems, Part I**
November 30, 1994

Abstract

This paper addresses the robust asymptotic and BIBO (bounded-input bounded-output) stability of a class of linear shift-variant multidimensional systems. Using a shift-invariant comparison system, necessary and sufficient conditions for the stability of the entire family of systems are derived.

1 Introduction

Results addressing the robust stability problem for 1-D discrete interval polynomials have generated interest in analogous results for the m-D case. Yet even in the work on shift-invariant m-D systems, only a few results address the $m > 2$ case [1, 2]. As for 1-D systems, conditions for the robust stability of shift-variant m-D systems are more restrictive than for the shift-invariant case. Some recent results concerning the robust stability of shift-variant m-D systems can be found in [3, 4, 5].

This work was supported in part by funds from the Office of Naval Research Grant #N00014-94-1-0387 and the SAE Foundation.

In [6], the authors investigated the BIBO stability of a class of shift-invariant 2-D systems having a nonessential singularity of the second kind (NSSK) on the distinguished boundary of the unit bidisk. This paper addresses what happens when we consider a shift-variant uncertain m-D system. In particular, we wish to determine whether a member of the family of shift-variant systems represented by a corresponding interval system may have a singularity on the distinguished boundary of the unit polydisk (T^m) and still be asymptotically and/or BIBO stable.

2 Notation

The work that follows requires some definitions and notation.

$\mathcal{N}_0^m$	The first m-D hyperquadrant.
$\overline{U}^m$	The closed unit polydisk: $\{(\underline{z}) : \lvert z_i\rvert \leq 1, i = 1, \cdots, m\}$
U^m	The open unit polydisk: $\{(\underline{z}) : \lvert z_i\rvert < 1, i = 1, \cdots, m\}$
T^m	The distinguished boundary of unit polydisk: $\{(\underline{z}) : \lvert z_i\rvert = 1, i = 1, \cdots, m\}$
$\underline{n}$, $\underline{i}$, $\underline{j}$	The spatial vectors $(n_1, \cdots, n_m)$, $(i_1, \cdots, i_m)$, and $(j_1, \cdots, j_m)$.
$y(\underline{n})$	The output of the m-D system.
$x(\underline{n})$	The input of the m-D system.
$a_{\underline{i}}(\underline{n})$	The shift-varying coefficient of a shifted output in a m-D difference equation. (For example, in a 2-D difference equation, $a_{(3,2)}(n_1, n_2)$ is the coefficient of $y(n_1 - 3, n_2 - 2)$.)
$b_{\underline{j}}(\underline{n})$	The shift-varying coefficient of a shifted input in a m-D difference equation.
N_j	The order of the m-D system in the n_j direction, $j = 1, \cdots, m$.
$\mathcal{I}$	$\{(i_1, \cdots, i_m) : 0 \leq i_k \leq N_k, k = 1, \ldots, m, \text{ and } (i_1, \cdots, i_m) \neq \underline{0}.\}$
$\mathcal{J}$	$\{(j_1, \cdots, j_m) : 0 \leq j_k \leq N_k, k = 1, \ldots, m, \text{ and } (j_1, \cdots, j_m) \neq \underline{0}.\}$

3 Problem Formulation

We consider m-D systems which are represented by the following m-D difference equation with shift-variant uncertain coefficients:

$$y(\underline{n}) = \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}) + \sum_{\underline{j} \in \mathcal{J}} b_{\underline{j}}(\underline{n}) x(\underline{n} - \underline{j}) \tag{1}$$

where

$$\begin{aligned} & a_{\underline{i}}(\underline{n}) \in [-a_{\underline{i}}^+, a_{\underline{i}}^+], \quad \forall \underline{i} \in \mathcal{I} \\ & b_{\underline{j}}(\underline{n}) \in [-b_{\underline{j}}^+, b_{\underline{j}}^+], \quad \forall \underline{j} \in \mathcal{J} \\ & a_{\underline{i}}^+, b_{\underline{j}}^+ \geq 0 \end{aligned} \tag{2}$$

We will also use the following shift-invariant majorant system, defined as follows:

$$y^+(\underline{n}) = \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}^+ y^+(\underline{n} - \underline{i}) + \sum_{\underline{j} \in \mathcal{J}} b_{\underline{j}}^+ x^+(\underline{n} - \underline{j}) \tag{3}$$

where

$$\begin{aligned} & x^+(\underline{i}) \geq |x(\underline{i})| \\ & y^+(\underline{i}) \geq |y(\underline{i})|, \quad \forall \underline{i} \in \mathcal{N}_0^m \end{aligned}$$

Note that for each $\underline{n} \in \mathcal{N}_0^m$, we may write a shift invariant difference equation with uncertain coefficients belonging to the intervals in (2):

$$y(\underline{n}) = \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}} y(\underline{n} - \underline{i}) + \sum_{\underline{j} \in \mathcal{J}} b_{\underline{j}} x(\underline{n} - \underline{j}) \tag{4}$$

where

$$\begin{aligned} & a_{\underline{i}} \in [-a_{\underline{i}}^+, a_{\underline{i}}^+], \quad \forall \underline{i} \in \mathcal{I} \\ & b_{\underline{j}} \in [-b_{\underline{j}}^+, b_{\underline{j}}^+], \quad \forall \underline{j} \in \mathcal{J} \\ & a_{\underline{i}}^+, b_{\underline{j}}^+ \geq 0 \end{aligned}$$

The set of systems represented by (4) has the following z-transform:

$$H_1(z_1, \cdots, z_m) = H_1(\underline{z}) = \frac{\sum_{\underline{j} \in \mathcal{J}} b_{\underline{j}} z_1^{j_1} \cdots z_m^{j_m}}{1 - \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}} z_1^{i_1} \cdots z_m^{i_m}} = \frac{N_1(\underline{z})}{D_1(\underline{z})} \tag{5}$$

We may also write the transfer function of the shift-invariant majorant system (3) as:

$$H_2(\underline{z}) = \frac{\sum_{\underline{j} \in \mathcal{J}} b_{\underline{j}}^+ z_1^{j_1} \cdots z_m^{j_m}}{1 - \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}^+ z_1^{i_1} \cdots z_m^{i_m}} = \frac{N_2(\underline{z})}{D_2(\underline{z})} \tag{6}$$

The results presented here make use of the following Lemma [7]:

Lemma 1 *The shift-variant uncertain system (1) is asymptotically (BIBO) stable if and only if the shift-invariant system (6) is asymptotically (BIBO) stable.*

We wish to know if it is possible for a family of shift-variant systems to have a shift-invariant member (4) with a singularity on T^m and still be asymptotically or BIBO stable. In addition, must the member producing such a singularity be the one with $a_{\underline{i}} = a_{\underline{i}}^+$ and $b_{\underline{j}} = b_{\underline{j}}^+$, $\forall \underline{i} \in \mathcal{I}$ and $\forall \underline{j} \in \mathcal{J}$?

4 Main Results

The following Corollary arises as an implication of Lemma 1:

Corollary 2 *If the shift-invariant interval transfer function (5) is asymptotically stable and has a member with a singularity on T^m, then the shift-variant family of systems (1) is asymptotically stable.*

Proof: If the interval system (5) is asymptotically stable, then the majorant system (6) is also asymptotically stable, since it is a member of (5). By Lemma 1, we can then conclude that the shift-variant system (1) is asymptotically stable. □

Remark: Given that the output mask of the majorant system (6) is at least two-dimensional, the following expression gives a necessary and sufficient condition for the asymptotic stability of (6)[3]:

$$\sum_{\underline{i}\in\mathcal{I}} a_{\underline{i}}^{+} \leq 1 \tag{7}$$

Theorem 3 *If the majorant system transfer function (6) has a singularity on T^m and no other singularities in $\overline{U}^m$, then the shift-variant family of systems (1) is BIBO unstable.*

Proof: Since Lemma 1 gives a necessary and sufficient condition for the BIBO stability of (1), we need only show that the majorant system (6) is BIBO unstable whenever it possesses a singularity on T^m.

An m-D transfer function cannot be BIBO-stable if it has singularities on T^m unless the singularities are nonessential singularities of the second kind (NSSK's). Because the majorant system has a singularity on T^m and no other singularities in $\overline{U}^m$, we have from [3] the following condition on the coefficients of $D_2(\underline{z})$ in (6):

$$\sum_{\underline{i}\in\mathcal{I}} a_{\underline{i}}^{+} = 1 \tag{8}$$

Clearly, $D_2(\underline{z})$ will have a zero at $(z_1, \cdots, z_m) = (1, \cdots, 1)$. Since none of the $b_{\underline{j}}^{+}$ in (6) are negative, $N_2(\underline{z})$ cannot have a zero at $(1, \cdots, 1)$. Thus the singularity at $(1, \cdots, 1)$ is a singularity of the first kind, or a pole, and thus $H_2(\underline{z})$ is not BIBO stable. From Lemma 1, the shift-variant system (1) is also BIBO unstable. □

5 Conclusion

In this paper we used a shift-invariant majorant system to derive conditions for the robust stability of shift-variant m-D systems. In particular, we examined the case of systems whose equivalent shift-invariant interval system representation has a singularity on T^m and no other singularities in $\overline{U}^m$. The following remarks discuss implications of the results presented here.

- The stability of the majorant system whose transfer function is given by (6) depends only on the denominator. The numerator cannot have a zero at $(z_1, \cdots, z_m) = (1, \cdots, 1)$, so the system cannot have an NSSK on T^m.

- If the interval system (5) contains a member with a singularity on T^m, then the shift-variant family of systems (1) may be robustly asymptotically stable (Corollary 2), but cannot be robustly BIBO stable (Theorem 3).

- The intervals to which the coefficients of (1) belong may be subintervals of $[-a_{\underline{i}}^+, a_{\underline{i}}^+]$ and/or $[-b_{\underline{j}}^+, b_{\underline{j}}^+]$. When such subintervals have the same upper limits as their corresponding intervals, the condition derived here is still necessary and sufficient for the stability of the shift-variant system. When the upper limit of any of the subintervals is not equal to the upper interval limit, we have a sufficient but not necessary condition for robust asymptotic stability.

- It is possible to construct stabilizing perturbations for a nominal BIBO stable system with an NSSK on T^m. Some directions of perturbation will result in a BIBO stable family of systems. This allows us to exploit the frequency response characteristics of low order filters with NSSK's on T^m, as long as we guarantee that perturbations from the nominal system are constrained to directions that result in robust stability.

References

[1] M. N. S. Swamy, L. M. Roytman, and E. I. Plotkin, "On stability properties of three- and higher dimensional linear shift-invariant digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 888–892, Sept. 1985.

[2] L. M. Roytman, N. Marinovic, and M. N. S. Swamy, "Sufficiency conditions for the stability of a class of 3-D functions with nonessential singularities of the second kind," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 1253–1255, Oct. 1987.

[3] P. H. Bauer, "Finite word-length effects in m-D digital filters with singularities on the stability boundary," *IEEE Trans. Signal Process.*, vol. 40, pp. 894–900, Apr. 1992.

[4] P. H. Bauer, "Robust stability of multi-dimensional (m-D) discrete interval systems and applications," *J. Circuits, Systems, and Computers*, vol. 1, no. 1, pp. 93–104, 1991.

[5] S. A. Yost and P. H. Bauer, "Robust stability of multidimensional difference equations with shift-variant coefficients," *Multidimens. Syst. Signal Proc.*, vol. 5, pp. 455–462, 1994.

[6] S. A. Yost, P. H. Bauer, and K. Balemarthy, "On the double bilinear transformation and nonessential singularities of the second kind at infinity," in *Proc. 1994 Int'l Symposium on Ckts. and Systems*, vol. 5, (London), pp. 137–140, June 1994.

[7] P. H. Bauer and E. I. Jury, "Boundary techniques for robust stability of time-variant discrete time systems," in *Control and Dynamic Systems - Advances in Theory and Applications* (C. T. Leondes, ed.), Academic Press, to appear 1995.

On the Double Bilinear Transformation and Nonessential Singularities of the Second Kind at Infinity

S. A. Yost
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

P. H. Bauer
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

K. Balemarthy
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana

Submitted to **Multidimensional Systems and Signal Processing**
December 1, 1994

Abstract

This paper addresses the BIBO (bounded-input bounded-output) stability of a class of discrete 2-D transfer functions in the presence of nonessential singularities of the second kind (NSSK's) on the unit bidisk. Conditions under which the double bilinear transformation (DBT) preserves stability are derived. The results presented here also extend the class of systems whose stability can be predicted. Use of the inverse DBT to produce a continuous equivalent of the discrete 2-D transfer function allows easy application of a continuous-domain equivalent of a criterion developed by Dautov. The necessary and sufficient condition for stability derived in this work provides a simple check for the class of systems under consideration. From this class of systems, it is also possible to construct stable pairs of mutually inverse transfer functions.

This work was supported in part by funds from the Office of Naval Research Grant #N00014-94-1-0387 and the SAE Foundation.

1 Introduction

In the study of multidimensional systems, the problem of assessing the BIBO (bounded-input bounded-output) stability of a 2-D system at a nonessential singularity of the second kind (NSSK) remains a salient research issue. For 2-D recursive digital filters, Goodman [1] showed that even if the transfer function has one or more NSSK's on the unit bidisk, it may still be stable. Since this important paper, other researchers have explored the issue, trying to learn more about the behavior of such systems in the presence of NSSK's.

Contributions in this area include the derivation of necessary and/or sufficient conditions for stability that improve upon previous results [2, 3, 4, 5], as well as the extension of Goodman's results to the n-D case, where $n > 2$ [6, 7]. Approaches to this problem vary widely. In [3], Alexander and Woods present a necessary condition based on the direction of tangents to the algebraic curve of the denominator polynomial at a zero on the unit bidisk. In [4], Roytman, Swamy, and Eichman use a resultant method to test for l_1- and l_2-stability. Recall the necessary and sufficient condition for l_1-stability:

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|h(m,n)| < \infty,$$

where $h(m,n)$ is the impulse response, or inverse $\mathcal{Z}$-transform of $H(z_1,z_2)$:

$$H(z_1,z_2) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}h(m,n)z_1^m z_2^n.$$

For l_2-stability we require the impulse response to be square-summable:

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|h(m,n)|^2 < \infty.$$

Dautov's approach [2] to determining the BIBO stability of a 2-D digital filter in the presence of an NSSK at $(\tilde{z}_1,\tilde{z}_2)$ on the distinguished boundary of the unit bidisk involves taking a limit as (z_1,z_2) approaches $(\tilde{z}_1,\tilde{z}_2)$. The path taken in this limit must be in the open unit bidisk. For an NSSK at $(1,1)$, consider the

path:

$$z_1 = 1 - y^2 + jy, \qquad |y| \leq 1$$

$$z_2 = z_1^*.$$

As $y \rightarrow 0$, (z_1, z_2) approaches $(1, 1)$. Dautov showed that if the limit of $H(z_1, z_2)$ exists along this path, then $H(z_1, z_2)$ is BIBO stable.

Reddy and Jury base their work [5] on Dautov's results. They translate Dautov's criterion to the continuous domain using the inverse double bilinear transformation (DBT). The DBT maps a continuous transfer function to a discrete function using

$$s_1 = \frac{1-z_1}{1+z_1} \quad s_2 = \frac{1-z_2}{1+z_2},$$

where without loss of generality, $T_1 = T_2 = 2$. The NSSK at $(z_1, z_2) = (1, 1)$ corresponds to an NSSK at $(s_1, s_2) = (0, 0)$. Reddy and Jury check the stability of the continuous function to infer the stability characteristics of the corresponding discrete transfer function.

The bilinear transformation is widely used in the design of 1-D recursive digital filters, and the DBT has been applied to 2-D digital filter design [8, 9, 10]. The 1-D bilinear transformation is known to preserve stability, but in the 2-D case, Goodman has shown in [11] that it is possible to apply the DBT to a stable continuous transfer function and obtain a discrete function that is unstable.

The authors of [12] formulate necessary and sufficient conditions for the BIBO stability of mutually inverse pairs of 2-D transfer functions having NSSK's on T^2. These conditions are based on the order of numerator and denominator zeros on T^2.

This paper presents a result regarding the stability of a class of discrete transfer functions whose inverse DBT has a particular form. We then give a sufficient condition for the BIBO stability of a class of 2-D continuous transfer functions that have NSSK's at $(s_1, s_2) = (\infty, \infty)$. This will allow us to gain insight into

the effect of the DBT on stability; i.e., the conditions under which a continuous system is stable and the corresponding discrete system is unstable. Previous work on the stability preserving properties of the DBT can also be found in [13]. Finally, we show that a class of BIBO stable transfer functions with stable inverses can be constructed from the class of systems under consideration.

2 Notation and Problem Formulation

The work that follows requires the following notation:

$\overline{U}^2 \doteq \{(z_1, z_2) : |z_i| \leq 1, \quad i = 1, 2\}$ closed unit bidisk

$U^2 \doteq \{(z_1, z_2) : |z_i| < 1, \quad i = 1, 2\}$ open unit bidisk

$T^2 \doteq \{(z_1, z_2) : |z_i| = 1, \quad i = 1, 2\}$ distinguished boundary of unit bidisk

$\overline{A}^2 \doteq \{(s_1, s_2) : \mathrm{Re}(s_i) \geq 0, \quad |s_i| < \infty, \quad i = 1, 2\}$

$A^2 \doteq \{(s_1, s_2) : \mathrm{Re}(s_i) > 0, \quad |s_i| < \infty, \quad i = 1, 2\}$

$A_0^2 \doteq \{(s_1, s_2) : \mathrm{Re}(s_i) = 0, \quad |s_i| < \infty, \quad i = 1, 2\}$

We will examine a class of discrete systems, denoted by $H(z_1, z_2)$, whose DBT has the following form:

$$H(s_1, s_2) = \frac{N(s_1, s_2)}{D(s_1, s_2)} \cdot \frac{\prod_{i=1}^{K}(1 + a_i s_1 + b_i s_2)}{\prod_{i=1}^{N}(1 + c_i s_1 + d_i s_2)} \tag{1}$$

which corresponds to the following class of discrete 2-D systems:

$$H_D(z_1, z_2) = \frac{N_D(z_1, z_2)}{D_D(z_1, z_2)} \cdot \frac{\prod_{i=1}^{K}[(1 + a_i + b_i) + (1 - a_i + b_i)z_1 + (1 + a_i - b_i)z_2 + (1 - a_i - b_i)z_1 z_2]}{\prod_{i=1}^{N}[(1 + c_i + d_i) + (1 - c_i + d_i)z_1 + (1 + c_i - d_i)z_2 + (1 - c_i - d_i)z_1 z_2]} \tag{2}$$

Note that the discrete system has an NSSK at $(z_1, z_2) = (-1, -1)$. Studying the asymptotic behavior of

$H(s_1, s_2)$ as s_1 and s_2 approach ∞ will tell us whether or not the system is stable at the NSSK. Note that this approach differs from that of Reddy and Jury [5] in that while they examine NSSK's at $(z_1, z_2) = (1, 1)$, we study the behavior of the transfer function at a different location on T^2: $(-1, -1)$. We do this because the path for $(s_1, s_2) \rightarrow (\infty, \infty)$ in the limit is easier to work with than the path for $(s_1, s_2) \rightarrow (0, 0)$.

We impose the following conditions on $H(s_1, s_2)$:

- The DBT of $\frac{N(s_1,s_2)}{D(s_1,s_2)}$ must be BIBO stable.
- $\frac{N(s_1,s_2)}{D(s_1,s_2)}$ does not contain any factors of the form $(1 + as_1 + bs_2)$, and it has no NSSK's at (∞, ∞).
- c_i and $d_i > 0$.
- $N \geq K$.

The presentation of the result requires the following definitions:

- Define the 1-D degree of $N(s, s)$ as α.
- Define the 1-D degree of $D(s, s)$ as β.
- Define the transfer functions:

$$G_r(s_1, s_2) = \frac{\prod_{i=1}^{K_r}(1 + a_i s_1 + b_i s_2)}{\prod_{i=1}^{N_r}(1 + c_i s_1 + d_i s_2)} \tag{3}$$

where

$$\frac{a_i}{b_i} = \frac{c_i}{d_i} = r,$$

except if $K_r = 0$, in which case the numerator expression is simply equal to 1, or if $N_r = 0$, in which case the denominator expression is equal to 1.

- Define $G_{ro}(s_1, s_2)$ as the $G_r(s_1, s_2)$ with the highest degree difference:

$$N_{ro} - K_{ro} = \max_r [N_r - K_r].$$

Thus,

$$G_{ro}(s_1, s_2) = \frac{\prod_{i=1}^{K_{ro}} (1 + a_i s_1 + b_i s_2)}{\prod_{i=1}^{N_{ro}} (1 + c_i s_1 + d_i s_2)}$$

where

$$\frac{a_i}{b_i} = \frac{c_i}{d_i} = r_o$$

except if $K_{ro} = 0$, in which case the numerator expression is simply equal to 1. (If there is more than one such G_r, then any one of them can be taken to be G_{ro}.)

3 Main Results

Theorem 1 *The DBT of $H(s_1, s_2)$ in (1) is BIBO stable if and only if*

$$(\alpha - \beta) + (K - K_{ro}) - (N - N_{ro}) \leq -1. \tag{4}$$

Proof:(Sufficiency) After examining bounds on

$$\left. \frac{\prod_{i=1}^{K} (1 + a_i s_1 + b_i s_2)}{\prod_{i=1}^{N} (1 + c_i s_1 + d_i s_2)} \right|_{\substack{s_1 \to \infty \\ s_2 \to \infty}}$$

and

$$\left.\frac{N(s_1,s_2)}{D(s_1,s_2)}\right|_{\substack{s_1\to\infty \\ s_2\to\infty}}$$

we will show that the stated condition is sufficient for BIBO stability. The path to infinity chosen for s_1 and s_2 must lie in A^2. This corresponds to the path specified by Dautov to $(z_1, z_2) = (-1, -1)$ in the discrete domain.

Consider the expression:

$$\frac{\prod_{i=1}^{K}(1+a_is_1+b_is_2)}{\prod_{i=1}^{N}(1+c_is_1+d_is_2)} = \frac{\prod_{i=1}^{K_{ro}}(1+a_is_1+b_is_2)\prod_{j=K_{ro}+1}^{K}(1+\tilde{a}_js_1+\tilde{b}_js_2)}{\prod_{i=1}^{N_{ro}}(1+c_is_1+d_is_2)\prod_{j=N_{ro}+1}^{N}(1+\tilde{c}_js_1+\tilde{d}_js_2)}$$

where because of how we defined G_{ro},

$$\frac{a_i}{b_i} = \frac{c_i}{d_i} = r_o$$

The choice of path for s_1 and s_2 must coincide with the requirements for the application of Dautov's criterion; namely, $\mathrm{Re}(s_1, s_2) > 0$ and $(s_1, s_2) \to (\infty, \infty)$. Depending on this choice, each factor in G_{ro} will either go to ∞ or remain bounded. (Actually, the preceding is true for *all* factors, including those indexed by j.) The difference between the number of unbounded factors in the numerator and denominator determines the asymptotic behavior of the expression. Without loss of generality, we choose the path:

$$\begin{aligned} s_1 &= \sigma + j\Omega \\ s_2 &= r_o(\sigma - j\Omega) \end{aligned} \tag{5}$$

This choice causes all of the factors in G_{ro} (those indexed by i) to approach a finite value for $\sigma > 0$ as $\Omega \to \infty$. The factors indexed by j will go to ∞ in the limit. Because G_{ro} has the highest degree difference of all the G_r, the path given by (5) results in the slowest convergence of $H(s_1, s_2)$ (or fastest divergence.)

Examining the asymptotic behavior of the expression containing factors indexed by j results in the observation that as s_1 and s_2 approach ∞ by the chosen path, the entire expression can be asymptotically bounded from above by $C_1 \cdot \Omega^{(K-K_{ro})-(N-N_{ro})}$, where C_1 is a finite constant.

Now we must derive an expression which describes the asymptotic behavior of $\frac{N(s_1,s_2)}{D(s_1,s_2)}$ as s_1 and s_2 approach ∞. First, note that because this expression has no NSSK's at (∞,∞), we can simplify the analysis by studying $\frac{N(\Omega,\Omega)}{D(\Omega,\Omega)}$ for $\Omega \rightarrow \infty$. If the degree of $N(\Omega,\Omega)$ is α and the degree of $D(\Omega,\Omega)$ is β, then the expression behaves asymptotically as $C_2 \cdot \Omega^{(\alpha-\beta)}$, where C_2 is a finite constant.

Combining the derived bounds as a product, we obtain an upper bound B_u for the $H(s_1,s_2)$ given by (1) for the critical path in (5):

$$B_u = C_1 C_2 \cdot \Omega^{(\alpha-\beta)+(K-K_{ro})-(N-N_{ro})} \tag{6}$$

If condition (4) is satisfied, then as $\Omega \rightarrow \infty$, $H(s_1,s_2) \rightarrow 0$ independent of the path in A^2, which implies BIBO stability of the DBT of $H(s_1,s_2)$.

(Necessity) For this part of the proof, we will show that for

(a) $(\alpha-\beta)+(K-K_{ro})-(N-N_{ro}) > 0$

(b) $(\alpha-\beta)+(K-K_{ro})-(N-N_{ro}) = 0,$

we will not obtain a unique limit of $H(s_1,s_2)$ at (∞,∞).

(a) Recall the upper bound on $H(s_1,s_2)$ derived as (6):

$$B_u = C_1 C_2 \cdot \Omega^{(\alpha-\beta)+(K-K_{ro})-(N-N_{ro})}$$

Since $H(s_1, s_2)$ actually grows with $\Omega^{(\alpha-\beta)+(K-K_{ro})-(N-N_{ro})}$, we can formulate a lower bound as:

$$B_l = D \cdot \Omega^{(\alpha-\beta)+(K-K_{ro})-(N-N_{ro})}$$

Clearly, if the exponent of Ω in (6) is positive, then as $\Omega \rightarrow \infty$ along the path in (5), $H(s_1, s_2)$ grows without bound. According to Dautov's criterion, if this limit does not exist, the transfer function is unstable.

(b) Now the exponent of Ω is zero, and $H(s_1, s_2)$ is bounded as $\Omega \rightarrow \infty$. It is trivial to show that in this case, the bound depends on the path taken; i.e., the limit at (∞, ∞) does not exist. □

Theorem 2 *The continuous transfer function:*

$$H(s_1, s_2) = \frac{N(s_1, s_2)}{D(s_1, s_2)} \cdot \prod_{r=1}^{R} G_r(s_1, s_2) \tag{7}$$

is BIBO stable if

- *for each* $G_r(s_1, s_2)$, *we have* $K_r \leq N_r$, *and* $c_i, d_i > 0$, *and*
- $\frac{N(s_1,s_2)}{D(s_1,s_2)}$ *is BIBO stable.*

Proof: Since we know that the cascade of BIBO stable transfer functions results in a BIBO stable system, showing that each of the $G_r(s_1, s_2)$ is stable will suffice to show the stability of $H(s_1, s_2)$.

First, we rewrite (3) as:

$$G_r(s_1, s_2) = \prod_{i=1}^{K_r} \left(\frac{1 + a_i s_1 + b_i s_2}{1 + c_i s_1 + d_i s_2} \right) \cdot \prod_{j=K_r+1}^{N_r} \left(\frac{1}{1 + c_j s_1 + d_j s_2} \right),$$

where

$$\frac{c_j}{d_j} = \frac{c_i}{d_i}.$$

Note that each of the factors indexed by j is a BIBO stable transfer function [13]. Therefore, the cascade of the $(N_r - K_r)$ BIBO stable transfer functions indexed by j is also BIBO stable.

Rearranging one of the factors indexed by i, and using the fact that $\frac{a_i}{b_i} = \frac{c_i}{d_i}$, we obtain:

$$\frac{1 + a_i s_1 + b_i s_2}{1 + c_i s_1 + d_i s_2} = \frac{a_i}{c_i} \cdot \left[1 + \frac{\frac{c_i}{a_i} - 1}{1 + c_i s_1 + d_i s_2}\right]$$

Therefore, each of the transfer functions indexed by i is BIBO stable, and we can infer the stability of the cascade of the K_r BIBO stable transfer functions indexed by i.

Using the same cascade argument, we can now make the statement that because each G_r consists of a cascade of BIBO stable transfer functions, each G_r can be said to be BIBO stable. And finally, since the product of the G_r is a BIBO stable transfer function, and $\frac{N(s_1,s_2)}{D(s_1,s_2)}$ is BIBO stable, then $H(s_1, s_2)$ is also stable. □

Corollary 1 *The double bilinear transformation fails to preserve stability when a continuous transfer function of the form given in (7) satisfies Theorem 2 and:*

$$N_{ro} - K_{ro} > -1 + (\beta - \alpha) + (N - K)$$

Proof: The proof follows directly from Theorem 1. If a BIBO stable continuous transfer function satisfies the above condition, then it *fails* to satisfy (4). □

This corollary helps us to identify classes of systems for which the DBT does not preserve stability. From Theorem 1, the inverse DBT of a BIBO stable discrete transfer function must satisfy:

$$N_{ro} - K_{ro} \leq -1 + (\beta - \alpha) + (N - K).$$

It is possible to identify some special cases that illustrate the instability of discrete systems obtained from stable continuous transfer functions using the DBT. First, consider an $H(s_1, s_2)$ for which $N(s_1, s_2)$ and $D(s_1, s_2)$ have the same degree, and for which in each G_r, $(r \neq r_0)$, $N_r = K_r$. Another class of systems addressed by Corollary 1 is the class of systems which consists of $N(s_1, s_2)/D(s_1, s_2)$ and a single G_r. As long as $K_r \leq N_r$ and $\beta \geq \alpha$, the continuous system is BIBO stable. For both classes of functions, as long as $\beta = \alpha$, the DBT fails to preserve stability. Summarizing these special cases, we have:

- $N_r = K_r$, $r \neq r_0$, $\alpha = \beta \Rightarrow (N - K) = (N_{ro} - K_{ro}) \Rightarrow$ DBT fails.
- $N_{ro} = N$, $K_{ro} = K \Rightarrow (N - K) = (N_{ro} - K_{ro}) \Rightarrow 0 \leq -1 + (\beta - \alpha) \Rightarrow$ DBT fails when $(\beta - \alpha) = 0$.

Note that for the above special cases, the transfer function $\frac{N(s_1,s_2)}{D(s_1,s_2)}$ can stabilize the overall transfer function as long as $\frac{N(s,s)}{D(s,s)}$ possesses one or more zeros at infinity; i.e., $(\beta - \alpha) > 0$.

Corollary 1 can be applied to the example given by Goodman in [11]:

$$H_1(s_1, s_2) = \frac{1}{1 + s_1 + s_2}.$$

According to Theorem 2, this function is BIBO stable, but according to Corollary 1, its DBT is unstable.

The DBT of the inverse of a BIBO stable member of the class of systems described by (1) cannot be BIBO stable because the limit of such a $H(s_1, s_2)$ as $(s_1, s_2) \rightarrow (\infty, \infty)$ is zero. This makes the corresponding limit of its inverse nonexistent. Thus from Dautov's criterion, this inverse is not BIBO stable.

However, by adding a non-zero constant C to an $H(s_1, s_2)$ whose DBT is BIBO stable we obtain:

$$\tilde{H}(s_1, s_2) = C + H(s_1, s_2)$$

The DBT of this modified transfer function is also BIBO stable. In particular, examining the behavior of the system at the NSSK at infinity, the limit of $\tilde{H}(s_1, s_2)$ exists and is equal to C. Furthermore, its inverse has a unique limit equal to $1/C$.

Note: $\tilde{H}(s_1, s_2)$ can be written as follows:

$$\tilde{H}(s_1, s_2) = \frac{C \cdot D(s_1, s_2) \prod_{i=1}^{N} (1 + c_i s_1 + d_i s_2) + N(s_1, s_2) \prod_{i=1}^{K} (1 + a_i s_1 + b_i s_2)}{D(s_1, s_2) \prod_{i=1}^{N} (1 + c_i s_1 + d_i s_2)} \tag{8}$$

If the numerator of the DBT of $\tilde{H}(s_1, s_2)$ has no zeros in $\overline{U}^2$ except for the NSSK at $(-1, -1)$, then the DBT of $\tilde{H}(s_1, s_2)$ and its inverse are a BIBO stable mutually inverse pair of 2-D discrete transfer functions. Existing tests for zero exclusion from $\overline{U}^2$ may be used to check the numerator of the DBT of $\tilde{H}(s_1, s_2)$.

4 Conclusion

In this paper, we have examined the BIBO stability of a class of 2-D continuous transfer functions and the stability of its discrete equivalent using the double bilinear transformation. The condition given in Theorem 1 allows the surprising result that the cascade of two or more BIBO *unstable* discrete transfer functions may actually be stable. For example, consider

$$H_2(s_1, s_2) = \left(\frac{1}{1 + c_1 s_1 + d_1 s_2} \right) \left(\frac{1}{1 + c_2 s_1 + d_2 s_2} \right),$$

where $\frac{c_1}{d_1} \neq \frac{c_2}{d_2}$. The DBT of $H_2(s_1, s_2)$ is stable according to Theorem 1, although the DBT of each factor of $H_2(s_1, s_2)$ is unstable. Comparing the sufficient condition for stability of the continuous system with the necessary and sufficient condition for stability of its DBT reveals the conditions under which, for the class of systems considered, the DBT fails to preserve stability.

We have also shown that it is possible to generate BIBO stable mutually inverse transfer functions from the class of 2-D continuous transfer functions under consideration.

References

[1] D. Goodman, "Some stability properties of two-dimensional linear shift-invariant digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 201–208, Apr. 1977.

[2] S. A. Dautov, "On absolute convergence of the series of Taylor coefficients of a rational function of two variables: Stability of two-dimensional recursive digital filters," *Sov. Math. Dokl.*, vol. 23, pp. 448–451, 1981.

[3] R. K. Alexander and J. W. Woods, "2-D digital filter stability in the presence of second kind nonessential singularities," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 604–612, Sept. 1982.

[4] L. M. Roytman, M. N. S. Swamy, and G. Eichman, "BIBO stability in the presence of nonessential singularities of the second kind in 2-D digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 60–72, Jan. 1987.

[5] H. C. Reddy and E. I. Jury, "Study of the BIBO stability of 2-D recursive digital filters in the presence of nonessential singularities of the second kind - analog approach," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 280–284, Mar. 1987.

[6] M. N. S. Swamy, L. M. Roytman, and E. I. Plotkin, "On stability properties of three- and higher dimensional linear shift-invariant digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 888–892, Sept. 1985.

[7] L. Wang and D. Xiyu, "Nonessential singularities of the second kind and stability for multidimensional digital filters," *Multidimens. Syst. Signal Proc.*, vol. 3, pp. 363–380, Oct. 1992.

[8] J. M. Costa and A. N. Venetsanopolous, "Design of circularly symmetric two-dimensional recursive filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-22, pp. 432–443, Dec. 1974.

[9] S. Chakrabarti, B. B. Battacharyya, and M. N. S. Swamy, "Approximation of two-variable specifications in the analog domain," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 378–388, July 1977.

[10] S. Chakrabarti and S. K. Mitra, "Design of two-dimensional filters via spectral transformation," *Proc. IEEE*, vol. 65, pp. 905–914, June 1977.

[11] D. Goodman, "Some difficulties with the double bilinear transformation in 2-D recursive filter design," *Proc. IEEE*, vol. 66, pp. 796–797, July 1978.

[12] M. N. S. Swamy and L. M. Roytman, "Stability of mutually inverse rational 2-d digital transfer functions," in *Proc. 1994 Int'l Symposium on Ckts. and Systems*, vol. 2, (London), pp. 597–600, June 1994.

[13] E. I. Jury and P. Bauer, "On the stability of 2-D continuous systems," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 1487–1500, Dec. 1988.

Asymptotic Stability of Linear Shift-Variant Difference Equations with Diamond-Shaped Uncertainties

S. A. Yost
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana
(219) 631-7850
yost.2@nd.edu

P. H. Bauer
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana
(219) 631-8015
pbauer@mars.ee.nd.edu

ABSTRACT

This paper addresses the asymptotic stability of linear shift-variant difference equations whose coefficients are uncertain in an m-dimensional hyperdiamond. The approach used here allows the construction of regions in the coefficient space guaranteeing asymptotic stability that extend beyond the region specified by existing results.

I. Introduction

The pioneering work of Kharitonov [1] concerning the Hurwitz stability of interval polynomials has generated considerable interest in finding ways to apply these results to representations of discrete time systems. The results described in [2–6] address the Schur stability of interval polynomials, which may be used to represent linear shift-invariant discrete time systems. The authors of [7] consider the Hurwitz stability of polynomials with the coefficients in a diamond. Among the reasons given for examining the diamond-shaped uncertainty structure (for both continuous and discrete systems) are the restrictiveness of independent coefficient variations and the possibility of generalizing results for the hyperdiamond to wider classes of regions. Other results obtained using a diamond-shaped uncertainty structure include those found in [8–11].

In the case of linear difference equations with shift-variant uncertain coefficients, we cannot generally rely on the frequency domain techniques which work for time-invariant systems. Here, the Schur stability of an interval polynomial or matrix does not guarantee the stability of a time-variant system whose parameters are constrained by the interval system. Some recent results on the robust stability of discrete time-variant systems can be found in [12–16]. Most of these results use state space representations to describe the system.

This work was supported in part by funds from the Office of Naval Research Grant #N00014-94-1-0387 and the SAE Foundation.

In [15], the authors examine linear shift-variant difference equations whose uncertain coefficients are constrained to lie in a hyperdiamond that is symmetric with respect to the origin in the coefficient space. They show that the unit hyperdiamond is the largest such region symmetric with respect to the origin that guarantees asymptotic stability. By reformulating the difference equation as a Schur stable nominal polynomial driven by coefficient perturbations about the nominal coeffcients (rather than with respect to the origin), we will generalize the results in [15].

II. Problem Formulation

Consider the m-th order difference equation with shift-variant uncertain coefficients:

$$y(n) = a_1(n)y(n-1) + \cdots + a_m(n)y(n-m), \quad (1)$$

where $a_i(n) = a_i^0 + \Delta a_i(n)$. This equation may also be written as:

$$y(n) = a_1^0 y(n-1) + \cdots + a_m^0 y(n-m) + e(n-1), \quad (2)$$

where

$$e(n-1) = \Delta a_1(n)y(n-1) + \cdots + \Delta a_m(n)y(n-m). \quad (3)$$

The uncertain $\Delta a_i(n)$ vary inside the hyperdiamond:

$$\sum_{i=1}^{m} |\Delta a_i(n)| = S \quad (4)$$

The nominal coefficients a_i^0 are chosen such that the coefficient vector $(a_1^0, \cdots, a_m^0)$ describes a Schur polynomial. If we define $h(n)$ to be the impulse response of (2), then define γ, a positive real number, as follows:

$$\gamma = \frac{1}{\sum_{n=0}^{\infty} |h(n)|}, \quad (5)$$

III. Main Result

Theorem 1 *The shift-variant system described by (2), (3), and (4)is asymptotically stable if*

$$S < \gamma \tag{6}$$

Proof:
The system output $y(n)$ may be expressed as the sum of two response components: the response due to arbitrary initial conditions, $y_I(n)$, and the response due to the input, $y_e(n)$. Thus, (2) can be expressed as:

$$y(n) = y_I(n) + y_e(n).$$

Suppose we know that for $n \leq N_0$, $|y(n)| \leq B_y^{(0)}$. We want to show that with increasing n, we can obtain successively smaller bounds on the output $|y(n)|$; i.e.,

$$\begin{array}{lll} |y(n)| \leq B_y^{(0)}, & \forall n \geq 0 \\ |y(n)| \leq B_y^{(1)} < B_y^{(0)}, & \forall n \geq N_0 \\ |y(n)| \leq B_y^{(i)} < B_y^{(i-1)}, & \forall n \geq N_{i-1} \end{array}$$

so that $B_y^{(i)} \to 0$ as $i \to \infty$, and thus, $|y(n)| \to 0$ as $n \to \infty$.

Step 1. Derive the following upper bound on $|e(n-1)|$:

$$|e(n-1)| = |\Delta a_1(n) y(n-1) + \cdots + \Delta a_m(n) y(n-m)|.$$

Since for $n \leq N_0$, we know that $|y(n)| \leq B_y^{(0)}$, we can write:

$$|e(n-1)| \leq (|\Delta a_1(n)| + \cdots + |\Delta a_m(n)|) B_y^{(0)}$$

$$|e(n-1)| \leq S \cdot B_y^{(0)}, \qquad n \leq N_0 + 1$$

$$B_e^{(0)} = S \cdot B_y^{(0)}, \qquad n \leq N_0 + 1 \tag{7}$$

Step 2. We obtain a bound for $|y(N_0+1)|$ using BIBO property to find a bound on the forced response:

$$|y_e(N_0+1)| = \left| \sum_{k=0}^{N_0+1} h(k) e(N_0 - k) \right|$$

$$|y_e(N_0+1)| < B_e^{(0)} \cdot \sum_{k=0}^{\infty} |h(k)| \tag{8}$$

Using (7) for $B_e^{(0)}$ in (8), we obtain the following upper bound for $|y_e(N_0+1)|$:

$$B_{y_e}^{(1)} = B_y^{(0)} \cdot S \cdot \sum_{k=0}^{\infty} |h(k)|. \tag{9}$$

Note that because our choice of nominal coefficients describes a Schur polynomial, the portion of the response due to initial conditions approaches zero asymptotically as n approaches infinity. More precisely, given $\epsilon_I > 0$, there exists $N_0 > 0$ such that for $n \geq N_0$, $|y_I(n)| < \epsilon_I$, or

$$|y(n)| < |y_e(n)| + \epsilon_I \tag{10}$$

We use this fact to write the bound on the total response, $|y(N_0+1)|$:

$$B_y^{(1)} = B_{y_e}^{(1)} + \epsilon_I \tag{11}$$

Step 3. Show that $B_y^{(1)} < B_y^{(0)}$.
Substituting (9) for $B_{y_e}^{(1)}$ in (11), we have:

$$B_y^{(1)} = B_y^{(0)} \cdot S \cdot \sum_{k=0}^{\infty} |h(k)| + \epsilon_I.$$

Now we solve the following inequality for S:

$$B_y^{(0)} \cdot S \cdot \sum_{k=0}^{\infty} |h(k)| + \epsilon_I < B_y^{(0)} \tag{12}$$

This provides a condition for which we attain a reduction in bound:

$$S < \frac{1}{\sum_{k=0}^{\infty} |h(k)|} - \frac{\epsilon_I}{B_y^{(0)} \sum_{k=0}^{\infty} |h(k)|}, \tag{13}$$

Now since we know from (6) that $S < \gamma$, we may write:

$$S = \gamma - \mu, \qquad 0 < \mu < \gamma \tag{14}$$

Using (5), we may rewrite (13) as:

$$S < \gamma - \frac{\epsilon_I}{B_y^{(0)} \sum_{k=0}^{\infty} |h(k)|}, \tag{15}$$

and we can obtain a reduced bound, since given μ, we can always find a sufficiently small ϵ_I such that

$$\frac{\epsilon_I}{B_y^{(0)} \sum_{k=0}^{\infty} |h(k)|} < \mu, \tag{16}$$

or

$$\epsilon_I < \frac{\mu}{\gamma} \cdot B_y^{(0)}$$

Step 4. Show that this reduced bound $B_y^{(1)}$ is valid for all $n \geq N_0 + 1$.
Consider $n = N_1 + 2$. Using the procedure of Step 1, we

can write:

$$|e(N_0+1)| \le S \cdot \max\{|y(N_0+1)|, \cdots, |y(N_0+2-m)|\}$$

$$|e(N_0+1)| \le S \cdot B_y^{(0)} \tag{17}$$

$|e(N_0+1)|$ has the same bound as $|e(N_0)|$, so the BIBO property tells us that the bound on $|y(N_0+2)|$ is the same as on $|y(N_0+1)|$; i.e., $B_y^{(1)}$.
If we consider any $n > N_0+1$, we have:

$$|e(n-1)| \le S \cdot \max_{1\le i\le m}\{|y(n-i)|\}$$

Since all of the $|y(n-i)|$ are bounded by either $B_y^{(0)}$ or $B_y^{(1)}$, we have:

$$|e(n-1)| \le S \cdot B_y^{(0)}, \qquad n \ge N_0+1, \tag{18}$$

and therefore, we obtain:

$$|y(n)| \le B_y^{(1)}, \qquad n \ge N_0+1, \tag{19}$$

Note that since $B_y^{(1)} < B_y^{(0)}$, we have (for $n \ge N_0+1$):

$$|y(n)| \le B_y^{(1)} = \alpha B_y^{(0)}, \qquad \alpha < 1, \tag{20}$$

Step 5. Because we model the system as a shift-variant nominal system with a time-variant driving input, we can examine the system response starting from any time instant. Here, we examine the response as if the system "started" at $n = N_0+m+1$. Note that in this case, the system has non-zero initial conditions bounded by $B_y^{(1)}$.

We may again view the system response as comprised of two parts:

$$y(n) = y_I(n) + y_e(n)$$

We again wait long enough for $|y_I(n)|$ to become sufficiently small. Or, given $\epsilon_I > 0$, there exists $N_1 > 0$ such that for $n \ge N_1$, $|y_I(n)| < \epsilon_I$, or

$$|y(n)| < |y_e(n)| + \epsilon_I \tag{21}$$

Step 6. Find new bound on $|e(n-1)|$ for $N_0+m+1 \le n \le N_1+1$; i.e.:

$$|e(n-1)| = |\Delta a_1(n)y(n-1) + \cdots + \Delta a_m(n)y(n-m)|.$$

Note that for $N_0+m+1 \le n \le N_1+1$, we know that

$$\max_{1\le i\le m}\{|y(n-i)|\} \le B_y^{(1)},$$

and

$$|e(n-1)| \le S \cdot B_y^{(1)}, \qquad N_0+m+1 \le n \le N_1+1$$

We have now a new bound on $|e(n-1)|$ that is valid from the time we consider the restarting of the system, up to and including $n = N_1+1$:

$$B_e^{(1)} = S \cdot B_y^{(1)} \tag{22}$$

Step 7. We may again use the BIBO property to find a bound on $|y(N_1+1)|$.

$$\begin{aligned} |y_e(N_1+1)| &\le \sum_{k=0}^{N_1-N_0-m} |h(k)||e(N_1-k)| \\ &\le B_e^{(1)} \cdot \sum_{k=0}^{N_1-N_0-m} |h(k)| < B_e^{(1)} \cdot \sum_{k=0}^{\infty} |h(k)| \end{aligned}$$

$$B_{y_e}^{(2)} = B_e^{(1)} \cdot \sum_{k=0}^{\infty} |h(k)| \tag{23}$$

Using (22) for $B_e^{(1)}$ in (23), we obtain:

$$B_{y_e}^{(2)} = B_y^{(1)} \cdot S \cdot \sum_{k=0}^{\infty} |h(k)| \tag{24}$$

To obtain a bound on the total response, we write:

$$B_y^{(2)} = B_{y_e}^{(2)} + \epsilon_I, \tag{25}$$

since $|y_I(n)| < \epsilon_I$ for $n \ge N_1$.

Step 8. Using the same argument as in Step 3, we can show that $B_y^{(2)} < B_y^{(1)}$, or that we can write

$$B_y^{(2)} = \alpha \cdot B_y^{(1)}. \tag{26}$$

And as in Step 4, $B_y^{(2)}$ is a valid bound on $|y(n)|$ for all $n \ge N_1+1$.

Step 9. Note that by the appropriate choice of the ϵ_I in (10) and in (21), we can write:

$$B_y^{(2)} = \alpha \cdot B_y^{(1)} = \alpha^2 \cdot B_y^{(0)}, \qquad 0 < \alpha < 1.$$

Step 10. By induction, using Steps 5-8 repeatedly, we find that $B_y^{(i)} = \alpha^i \cdot B_y^{(0)}$ is a valid bound on $|y(n)|$ that can be constructed for the system for $n \ge N_{i-1}$.
Clearly, as $i \to \infty$, $B_y^{(i)} \to 0$. Thus, as $n \to \infty$, $|y(n)| \to 0$.

IV. Examples

Example 1. In (2), let $a_i^0 = 0$, $i = 1, \ldots, m$. Since $h(n) = \delta(n)$ for such a system, $\gamma = 1$, and we have the unit hyperdiamond of [15] as the region which guarantees asymptotic stability.

Example 2. Consider a second order nominal system whose characteristic equation has real or imaginary roots. It can be shown that γ is equal to the size of the largest diamond centered at (a_1^0, a_2^0) that fits inside the triangular region of Schur stability. If the characteristic equation of the nominal system has complex roots, the value of γ may be slightly less than the size of the largest diamond that fits in the triangle.

V. Conclusion

By using a nominal Schur stable system driven by perturbations to model an uncertain shift-variant discrete system, we have established a new method for the construction of diamond shaped uncertainties around the nominal system which preserve stability of the resulting shift-variant system. The condition derived here is sufficient for asymptotic stability of the class of systems considered, and we expect that further work will show that the condition is also necessary for stability; that is, that the size of the hyperdiamond describing the perturbations about the nominal coefficient vector must not exceed the size specified by Theorem 1. Two observations support this prediction: (a) when we allow the nominal coefficient vector to be the origin, we obtain the same result as in [15], and (b) in some cases the hyperdiamond of the result presented here extends to the boundary of the region of Schur stability.

References

[1] V. L. Kharitonov, "Asymptotic stability of an equilibrium position of a family of systems of linear differential equations," *Differential'nye Uravnenia*, vol. 14, pp. 2086–2088, 1978.

[2] C. V. Hollot and A. C. Bartlett, "Some discrete-time counterparts to Kharitonov's stability criterion for uncertain systems," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 355–356, Apr. 1986.

[3] F. Kraus, B. D. O. Anderson, and M. Mansour, "Robust Schur polynomial stability and Kharitonov's theorem," *Int. J. Contr.*, vol. 47, no. 5, pp. 1213–1225, 1988.

[4] F. Kraus, B. D. O. Anderson, E. I. Jury, and M. Mansour, "On the robustness of low-order Schur polynomials," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 570–577, May 1988.

[5] N. K. Bose, E. I. Jury, and E. Zeheb, "On robust Hurwitz and Schur polynomials," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 1166–1168, Dec. 1988.

[6] Y. K. Foo and Y. C. Soh, "Schur stability of interval polynomials," *IEEE Trans. Automat. Contr.*, vol. AC-38, pp. 943–946, June 1993.

[7] N. K. Bose and K. D. Kim, "Stability of a complex polynomial set with coefficients in a diamond and generalizations," *IEEE Trans. Circuits Syst.*, vol. CAS-36, pp. 1168–1174, Sept. 1989.

[8] J. Kogan, "Hurwitz stability of weighted diamond polynomials," *Syst. and Contr. Letters*, pp. 303–312, Apr. 1994.

[9] B. R. Barmish, R. Tempo, C. V. Hollot, and H. I. Kang, "An extreme point result of robust stability of a diamond of polynomials," *IEEE Trans. Automat. Contr.*, vol. AC-37, pp. 1460–1462, Sept. 1992.

[10] Y. K. Foo and Y. C. Soh, "Stability of a family of polynomials with coefficients bounded in a diamond," *IEEE Trans. Automat. Contr.*, vol. AC-36, pp. 1501–1502, June 1993.

[11] A. Katbab and E. I. Jury, "Robust Schur polynomial stability of a complex-coefficient polynomials set with coefficients in a diamond," *J. Franklin Inst.*, vol. 327, no. 5, pp. 687–698, 1990.

[12] P. H. Bauer and K. Premaratne, "Robust stability of time-variant interval matrices," in *Proc. 29th IEEE Conf. Decision Contr.*, (Honolulu, HI), pp. 434–435, Dec. 1990.

[13] S. R. Kolla, R. A. Yedavalli, and J. B. Farison, "Robust stability bounds of time-varying perturbations for state space models of linear discrete-time systems," *Int. J. Contr.*, vol. 50, no. 1, pp. 151–159, 1989.

[14] F. Mota, E. Kaszkurewicz, and A. Bhaya, "Robust stabilization of time-varying discrete interval systems," in *Proc. 31th IEEE Conf. Decision Contr.*, (Tucson, AZ), pp. 341–346, Dec. 1992.

[15] P. H. Bauer, M. Mansour, and J. Durán, "Stability of polynomials with time-variant coefficients," *IEEE Trans. Circuits Syst.*, vol. CAS-40, pp. 423–426, June 1993.

[16] P. H. Bauer, K. Premaratne, and J. Durán, "A necessary and sufficient condition for robust asymptotic stability of time-variant discrete systems," *IEEE Trans. Automat. Contr.*, vol. AC-38, pp. 1427–1430, Sept. 1993.

Robust Asymptotic Stability of 2-D Shift-Variant Discrete State-Space Systems

S. A. Yost and P. H. Bauer

Abstract

The results described in this paper provide conditions for the asymptotic stability of 2-D shift-variant uncertain systems expressed using the Roesser state-space description. A necessary and sufficient condition for the asymptotic stability of 1-D systems involves checking all products of extreme matrices. The same test is shown to apply to 2-D systems, although the corresponding stability condition is sufficient, but not necessary.

1 Introduction

The problem of the stability of discrete interval matrix systems has sparked considerable interest in recent years. Many of the early results in this area focused on linear shift-invariant systems [1, 2]. Because many discrete systems exhibit quantization and other nonlinearities,

The authors are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556-5637.

This work was supported in part by funds from the Office of Naval Research Grant #N00014-94-1-0387 and the SAE Foundation.

some of the effort concentrated on the study of linear shift-variant systems, which may be used to model some nonlinear systems [3–9].

The shift-variant problem requires an approach quite different from those used in the shift-invariant case. The Schur stability of a family of shift-invariant matrices does not necessarily imply the stability of a shift-variant counterpart. Until recently, the results obtained for shift-variant systems were either conservative sufficient conditions or necessary and sufficient conditions that applied only to system matrices of very particular structures.

In this paper, we will show that recent results addressing the robust stability of 1-D discrete shift-variant systems [9] can be extended to the 2-D case. The following section introduces the necessary notation and the problem formulation. Section 3 gives the main result; i.e., it shows how an important necessary and sufficient condition for the stability of 1-D shift-variant discrete systems can be adapted for the 2-D case as a slightly more conservative sufficient condition.

2 Notation and Problem Formulation

In this paper, we consider 2-D shift-variant uncertain systems described using the Roesser state-space representation:

$$\begin{aligned}\begin{bmatrix} \underline{x}^h(n_1+1,n_2) \\ \underline{x}^v(n_1,n_2+1) \end{bmatrix} &= A(n_1,n_2)\underline{x}(n_1,n_2) \\ &= \begin{bmatrix} A_{HH}(n_1,n_2) & A_{HV}(n_1,n_2) \\ A_{VH}(n_1,n_2) & A_{VV}(n_1,n_2) \end{bmatrix} \begin{bmatrix} \underline{x}^h(n_1,n_2) \\ \underline{x}^v(n_1,n_2) \end{bmatrix}. \end{aligned} \tag{1}$$

where $\underline{x}(n_1,n_2) \in \Re^{(H+V)}$ is the state vector in a 2-D system of order H in n_1, and V in n_2. This vector may also be written as:

$$\underline{x}(n_1,n_2) = \begin{bmatrix} \underline{x}^h(n_1,n_2) \\ \underline{x}^v(n_1,n_2) \end{bmatrix}$$

$A(n_1,n_2)$ is the uncertain system matrix for a shift-variant 2-D system, with $A_{HH}(n_1,n_2) \in \Re^{H\times H}$, $A_{HV}(n_1,n_2) \in \Re^{H\times V}$, $A_{VH}(n_1,n_2) \in \Re^{V\times H}$, and $A_{VV}(n_1,n_2) \in \Re^{V\times V}$. Note that we can describe the uncertainty structure of the 2-D system as $A(n_1,n_2) \in \mathbf{A}^I$, where $\mathbf{A}^I \in \Re^{(H+V)\times(H+V)}$ is the interval matrix which describes the set of matrices

$$\mathbf{A}^I = \{A \in \Re^{(H+V)\times(H+V)} : A = A_0 + \sum_{i=1}^{p} \lambda_i(n_1,n_2)A_i\}, \tag{2}$$

where $\lambda_i(n_1,n_2) \in [\underline{\lambda}_i, \overline{\lambda}_i]$, $i = 1,\cdots,p$; A_0, $A_i \in \Re^{(H+V)\times(H+V)}$, and p is any positive integer.

The results described in this paper require the following additional notation:

- The decomposition of $A(n_1, n_2)$ defined as follows:

$$J(n_1,n_2) = \begin{bmatrix} A_{HH}(n_1,n_2) & A_{HV}(n_1,n_2) \\ 0 & 0 \end{bmatrix} \qquad K(n_1,n_2) = \begin{bmatrix} 0 & 0 \\ A_{VH}(n_1,n_2) & A_{VV}(n_1,n_2) \end{bmatrix}$$

- The variable dimension supermatrix defined as follows:

$$A_{cc}(n) = \begin{bmatrix} J(n,0) & 0 & \cdots & 0 \\ K(n,0) & J(n-1,1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(0,n) \\ 0 & 0 & \cdots & K(0,n) \end{bmatrix} \tag{3}$$

  Note that $A_{cc}(n) \in \Re^{(n+2)(H+V)\times(n+1)(H+V)}$.

- Let $\lambda_i(n)$ be the $\lambda_i(n_1, n_2)$ in (2) corresponding to $A_{cc}(n)$ along the diagonal $n = n_1+n_2$.

  Note that because $A_{cc}(n)$ has $(n+1)\times(H+V)$ columns, there are $(n+1)p$ such $\lambda_i(n)$.

- Let $A^E_{i,cc} = A_{cc}(n)$ with $\lambda_i(n) = \underline{\lambda}_i$ or $\overline{\lambda}_i$.

- Let $\mathbf{A}^E_{cc}(n)$ be the set of all extreme $A_{cc}(n)$: $\mathbf{A}^E_{cc}(n) = \{A^E_{i,cc}\}$, $i = 1, \cdots, 2^{(n+1)p}$

- Let $\mathbf{P}^E_{k,cc}$ be the set of all length k products of extreme matrices:

$$\mathbf{P}^E_{k,cc} = \{P^E_{k,cc}(i_{k-1}, \cdots, i_0) = A^E_{i_{k-1},cc}(k-1)A^E_{i_{k-2},cc}(k-2)\cdots A^E_{i_0,cc}(0) :$$
$$i_\nu \in \{1, \cdots, 2^{(\nu+1)p}\}, \nu = 0, \ldots, k-1\}.$$

We wish to derive conditions concerning the asymptotic stability of the 2-D shift-variant system (1), with structured uncertainties as described in (2). We will extend the 1-D results

found in [9] to the 2-D case. We use the following definition for the asymptotic stability of a multidimensional system [10]:

Definition 1 *A m-D first hyperquadrant causal digital filter is said to be asymptotically stable under all finitely extended bounded input signals* $e(n_1, \cdots, n_m)$ *where*

$$|e(n_1, \cdots, n_m)| \leq M \text{ for } n_1 + \cdots + n_m \leq D$$

$$e(n_1, \cdots, n_m) = 0 \text{ for } n_1 + \cdots + n_m > D$$

$$n_\nu \geq 0, \quad \nu = 1, \cdots, m$$

and M *being a real bounded number,* D *being some positive integer, if all the outputs of the m-D digital filter asymptotically reach zero for* $(n_1 + \cdots + n_m) \to \infty$, $\underline{n} \in \mathcal{N}_0^m$. *(Note:* $\mathcal{N}_0^m$ *is the first m-D hyperquadrant.)*

This definition is less restrictive than definitions which assume zero input and finite initial conditions on the boundary of the first hyperquadrant. Here we may allow finitely extended bounded input signals to drive the system. Eventually the system will operate under zero-input conditions for $(n_1 + \cdots + n_m)$ large enough. We need this less restrictive definition to guarantee the generation of system initializations which cannot be reached using only initial conditions on the boundary of the first hyperquadrant.

The main result of this paper was proposed in a previous article [11], but it lacks a formal proof. The main contribution this paper makes is a rigorous proof of a theorem that allows

for a finite stability test. When we state the 2-D result of [11] in Theorem 1, we use a slightly different notation to conform to the subsequent proof. The appeal of a necessary and sufficient stability criterion that can be implemented as a finite test motivates this work.

3 Main Result

Theorem 1 *The 2-D system in (1) is asymptotically stable in the sense of Definition 1 if there exists a finite k such that:*

$$\max_{P^E_{k,cc}(i_{k-1},\cdots,i_0)\in \mathbf{P}^E_{k,cc}} \|P^E_{k,cc}(i_{k-1},\cdots,i_0)\|_1 \leq \gamma < 1 \tag{4}$$

Proof: Define the 1-D vector $\underline{\phi}(n)$ as follows:

$$\underline{\phi}(n) = [\underline{x}^h(n,0)^T, \underline{x}^v(n,0)^T, \underline{x}^h(n-1,1)^T, \ldots, \underline{x}^h(0,n)^T, \underline{x}^v(0,n)^T]^T \tag{5}$$

This vector contains the values of all of the state variables along the diagonal $n_1 + n_2 = n$. If $\underline{\phi}(n) \rightarrow \underline{0}$ as $n \rightarrow \infty$, then we have asymptotic stability.

To study the asymptotic behavior of $\underline{\phi}(n)$, we write:

$$\underline{\phi}(n+1) = A_{cc}(n)\underline{\phi}(n),$$

where $A_{cc}(n)$ is as shown in (3). We may also write:

$$\underline{\phi}(n+1) = A_{cc}(n)A_{cc}(n-1)\cdots A_{cc}(n-k)\underline{\phi}(n-k),$$

or in particular:

$$\underline{\phi}(n+1) = A_{cc}(n)A_{cc}(n-1)\cdots A_{cc}(0)\underline{\phi}(0).$$

Thus, we can write $\underline{\phi}(n+1)$ as follows:

$$\underline{\phi}(n+1) = \begin{bmatrix} J(n,0) & 0 & \cdots & 0 \\ K(n,0) & J(n-1,1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(0,n) \\ 0 & 0 & \cdots & K(0,n) \end{bmatrix} \begin{bmatrix} J(n-1,0) & \cdots & 0 \\ K(n-1,0) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(0,n-1) \end{bmatrix} \cdots \begin{bmatrix} J(0,0) \\ K(0,0) \end{bmatrix} \underline{\phi}(0)$$

To prove that the maximum 1-norm of the product matrix occurs for one of the extreme product matrices, we must show that each entry in the product matrix is a multilinear function of the $\lambda_i(n)$. Consider the form of $A_{cc}(n)$:

$$A_{cc}(n) = \begin{bmatrix} J(n,0) & 0 & \cdots & 0 \\ K(n,0) & J(n-1,1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(0,n) \\ 0 & 0 & \cdots & K(0,n) \end{bmatrix}$$

Note the location in the spatial plane of each entry of $A_{cc}(n)$. Each $J(n_1, n_2)$ and $K(n_1, n_2)$ lies on the diagonal $n_1 + n_2 = n$. Now consider $A_{cc}(n-1)$. For this supermatrix, each $J(n_1, n_2)$ and $K(n_1, n_2)$ lies on the diagonal $n_1 + n_2 = n - 1$. Thus, an entirely different set of points in the spatial plane corresponds to the entries in $A_{cc}(n-1)$. Now when we form the product $A_{cc}(n)A_{cc}(n-1)$, we obtain:

$$A_{cc}(n)A_{cc}(n-1) = \begin{bmatrix} J(n,0)J(n-1,0) & \cdots & 0 \\ K(n,0)J(n-1,0) + J(n-1,1)K(n-1,0) & \cdots & 0 \\ K(n-1,1)K(n-1,0) & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ 0 & \cdots & J(1,n-1)J(0,n-1) \\ 0 & \cdots & K(1,n-1)J(0,n-1) + J(0,n)K(0,n-1) \\ 0 & \cdots & K(0,n)K(0,n-1) \end{bmatrix}$$

Note that each entry in this product matrix is multilinear in the uncertain parameters of the 2-D system. Continuing to right multiply by $A_{cc}(n-2)$ and so on, we see that we are introducing parameters at points in the spatial plane distinct from those already part of the product, so the multilinearity of entries in the product matrix $A_{cc}(n)A_{cc}(n-1)\cdots A_{cc}(n-k)$ is preserved. This multilinearity of entries allows the assertion that the maximum 1-norm occurs at one of the extreme product matrices.

Note also that if we consider:

$$\underline{\phi}(M+m+1) = A_{cc}(M+m)\cdots A_{cc}(M)\underline{\phi}(M)$$

and

$$\underline{\phi}(m+1) = A_{cc}(m)\cdots A_{cc}(0)\underline{\phi}(0),$$

then

$$\max_{\lambda_i(n),n=M,\cdots,M+m} \|A_{cc}(M+m)\cdots A_{cc}(M)\|_1 = \max_{\lambda_i(n),n=0,\cdots,m} \|A_{cc}(m)\cdots A_{cc}(0)\|_1,$$

since each column of the product matrix has the same form. Because this maximum 1-norm is independent of M, we can let $M = 0$ without loss of generality. In general, $A_{cc}(m)A_{cc}(m-1)$ has $m\times(H+V)$ columns, each of which has the same form, independent of m. We can extend this argument to a longer product of supermatrices, $A_{cc}(m)A_{cc}(m-1)A_{cc}(m-2)$. Here, the form of the columns changes, but again, each column has the same form, independent of m. Thus, $\max\|A_{cc}(m)A_{cc}(m-1)A_{cc}(m-2)\|_1$ will be the same as $\max\|A_{cc}(2)A_{cc}(1)A_{cc}(0)\|_1$ for any integer $m \geq 2$. Because the norm is independent of m, finding a finite k for which $\max_{P^E_{k,cc}(i_{k-1},\cdots,i_0)\in\mathbf{P}^E_{k,cc}} \|P^E_{k,cc}(i_{k-1},\cdots,i_0)\|_1 \leq \gamma < 1$ suffices to verify the stability of (1). □

Remarks:

- The formulation of this 2-D result allows the application of the computer-aided test described in [9] to 2-D shift-variant systems.

- The algorithm for the 1-D case allows the use of essentially any induced matrix norm. When implemented for 2-D systems, the algorithm uses only the 1-norm. The ∞-norm

cannot be used, since the rows of the product matrix do not have the same form, and thus in general,

$$\max \|A_{cc}(M+m)\cdots A_{cc}(M)\|_\infty \neq \max \|A_{cc}(m)\cdots A_{cc}(0)\|_\infty.$$

- Unlike for the 1-D case, $\max_{P^E_{k,cc}(i_{k-1},\cdots,i_0)\in \mathbf{P}^E_{k,cc}} \|P^E_{k,cc}(i_{k-1},\cdots,i_0)\|_1 \leq \gamma < 1$ is not necessary for asymptotic stability, but it is not believed that this condition is very conservative. Recall that

$$\underline{\phi}(n+1) = \left(\prod_{k=0}^{n} A_{cc}(k)\right) \underline{\phi}(0).$$

Since as $n \to \infty$, the length of $\underline{\phi}$ grows without bound, each component of $\underline{\phi}$ may approach 0 without the sum of such components approaching 0. So $\|\underline{\phi}(n)\|_1 \to 0$ is sufficient, but not necessary for stability, and thus, $\max_{P^E_{k,cc}(i_{k-1},\cdots,i_0)\in \mathbf{P}^E_{k,cc}} \|P^E_{k,cc}(i_{k-1},\cdots,i_0)\|_1 \leq \gamma < 1$ is also sufficient but not necessary.

4 Conclusion

In this paper, we have presented a rigorous proof of a theorem which gives a sufficient condition for the asymptotic stability of 2-D linear, shift-variant discrete systems. This theorem extends previous results for 1-D systems, and it allows the implementation of an algorithm that will converge in a finite number of steps for almost any stable 2-D system. For

2-D systems, the computational complexity of the test is much higher than for 1-D systems. The results can be generalized to the m-D case, but the computational complexity increases even more than for the 2-D case.

The condition in the theorem is not necessary for the asymptotic stability of 2-D systems, but it is not believed that this condition is very conservative. Work to find a necessary and sufficient condition for instability is underway, since the theorem does not allow for the conclusion that a system is unstable.

References

[1] M. Mansour, "Robust stability of interval matrices," in *Proc. 28th IEEE Conf. Decision Contr.*, (Tampa, FL), pp. 46–51, Dec. 1989.

[2] B. R. Barmish, *New Tools for Robustness of Linear Systems.* New York: Macmillan Publishing Co., 1994.

[3] M. Mansour, "Sufficient conditions for the asymptotic stability of interval matrices," *Int. J. Contr.*, vol. 47, no. 1, pp. 1973–1974, 1988.

[4] S. R. Kolla, R. A. Yedavalli, and J. B. Farison, "Robust stability bounds of time-varying perturbations for state space models of linear discrete-time systems," *Int. J.*

Contr., vol. 50, no. 1, pp. 151–159, 1989.

[5] E. Yaz and X. Niu, "Stability robustness of linear discrete-time systems in the presence of uncertainties," *Int. J. Contr.*, vol. 50, no. 1, pp. 173–182, 1989.

[6] M. Mansour, "Simplified sufficient conditions for the asymptotic stability of interval matrices," *Int. J. Contr.*, vol. 50, no. 1, pp. 443–444, 1989.

[7] G. Mayer, "On the convergence of powers of interval matrices," *Linear Algebra Appl.*, vol. 58, p. 201, 1984.

[8] P. H. Bauer and K. Premaratne, "Robust stability of time-variant interval matrices," in *Proc. 29th IEEE Conf. Decision Contr.*, (Honolulu, HI), pp. 434–435, Dec. 1990.

[9] P. H. Bauer, K. Premaratne, and J. Durán, "A necessary and sufficient condition for robust asymptotic stability of time-variant discrete systems," *IEEE Trans. Automat. Contr.*, vol. AC-38, pp. 1427–1430, Sept. 1993.

[10] P. H. Bauer, "Finite word-length effects in m-D digital filters with singularities on the stability boundary," *IEEE Trans. Signal Process.*, vol. 40, pp. 894–900, Apr. 1992.

[11] P. H. Bauer and E. I. Jury, "A stability analysis of two-dimensional nonlinear digital state-space filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-38, pp. 1578–1586, Sept. 1990.

Multidimensional Systems and Signal Processing, 5, 455–462 (1994)

Robust Stability of Multidimensional Difference Equations with Shift-Variant Coefficients

S.A. YOST AND P.H. BAUER
Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556-5637

Abstract. This paper addresses the asymptotic stability of multidimensional systems represented by first hyperquadrant causal linear difference equations whose coefficients are shift-varying. The results extend previous 1-D results, and include the derivation of a fixed region of stability in the parameter space, as well as a sequence of shift-variant parameter regions. In the case of a fixed parameter region, the largest stable hyperdiamond centered at the origin will be obtained. For the shift-variant case, it will be shown that the instantaneous stable parameter region always includes this hyperdiamond.

Key Words: m-D stability, robustness, shift-variant systems, structured uncertainties, asymptotic stability

1. Introduction

Most of the recent results on the problem of robust stability of discrete systems apply to shift-invariant systems. A one-dimensional (1-D) difference equation that characterizes a shift-invariant discrete system is stable if and only if the corresponding polynomial in z has none of its zeros in the closed unit disk. Here, the z-transform is defined with respect to positive powers of z. For the m-D case, the polynomial in $z_1, \ldots, z_m$ is stable if there are no zeros in the closed unit polydisk, except possibly at a finite number of locations on the distinguished boundary of the unit polydisk [1].

The problem of robust stability under structured uncertainties has recently been addressed for the one-dimensional discrete-time case [2]–[6]. Unfortunately, the results obtained for time-invariant discrete-time polynomials are much more complicated than their continuous-time counterparts [7], since no simple vertex results are possible. Very recently, the problem of time-variant discrete-time systems was addressed in [8], [9]. Necessary and sufficient conditions for stability were derived for polytopic uncertainties in discrete-time state-space models [9]. These results also apply to time-variant interval polynomials. The work in [8] considers the problem of finding the largest parameter region containing the origin which ensures global asymptotic stability of the time-variant system. Other recent results on the stability of time-variant state matrices can be found in [10], [11].

Unfortunately, very little is known about robust stability of shift-invariant or shift-variant m-D systems with structured uncertainties. For the 2-D case, a 1-D stability robustness result was used in [12] to develop a stability test for 2-D shift-invariant state-space systems. Some other results, which can be considered an extension of the 1-D case addressed in

[13], can be found in [14]–[16]. Because these results allow only hypercuboidal types of uncertainties, they often lead to conservative sufficient conditions which are necessary only for special cases.

This paper attempts to reduce the conservativeness in the work of [13]–[16] by following an approach introduced for 1-D systems in [8], [9]. This leads to larger stable regions in the coefficient space. In particular, we show the existence of a fixed hyperdiamond in the coefficient space which guarantees global asymptotic stability of a shift-varying polynomial. Furthermore, the region constructed will be shown to be the largest such region to guarantee stability. Finally, the existence of a shift-variant region of stability with infinite volume in the parameter space will be shown. Obviously, the results obtained for the shift-variant case can also be used as sufficient conditions for the shift-invariant case. In some cases, the parameter regions for the shift-variant and shift-invariant cases are shown to be identical.

2. Notation and problem formulation

We require some definitions and notation.

q	shift operation (z in the shift-invariant case)
$\mathfrak{N}_0$	set of nonnegative integers
$\mathfrak{N}_0^m$	first m-D hyperquadrant
$\underline{n}$, $\underline{i}$	spatial vectors $(n_1, \ldots, n_m)$ and $(i_1, \ldots, i_m)$
$y(\underline{n})$	output of the m-D system
$a_{\underline{i}}(\underline{n})$	shift-varying coefficient of a shifted output in a m-D difference equation (for example, in a 2-D difference equation, $a_{(3,2)}(n_1, n_2)$ is the coefficient of $y(n_1 - 3, n_2 - 2)$)
$\underline{a}(\underline{n})$	ordered vector of the coefficients of a m-D difference equation
N_j, $j = 1, \ldots, m$	order of the m-D system in the n_j direction
$\mathfrak{A}$	fixed region in the coefficient space that guarantees asymptotic stability
$\mathfrak{A}(\underline{n})$	sequence of regions in the coefficient space that guarantees asymptotic stability
$\mathfrak{K}_K$	$\{\underline{n} : n_1, + \cdots + n_m = K\}$, $n_1, \ldots, n_m, K \in \mathfrak{N}_0$
$\tilde{\mathfrak{K}}_K$	$\{\underline{n} : n_1 + \cdots + n_m \leq K\}$
B	$\max_{\underline{n} \in \tilde{\mathfrak{K}}_K} \{\lvert y(\underline{n}) \rvert\}$
$\mathfrak{I}$	$\{(i_1, \ldots, i_m) : 0 \leq i_j \leq N_j, j = 1, \ldots, m, \text{ and } (i_1, \ldots, i_m) \neq 0\}$
γ	fixed real number such that $0 \leq \gamma < 1$

To formulate the problem, we consider the following class of m-D first hyperquadrant causal linear difference equations with shift-variant coefficients:

$$y(\underline{n}) = \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}), \tag{1}$$

where the uncertainty structure is described as follows:

$$\sum_{\underline{i} \in \mathcal{I}} |a_{\underline{i}}(\underline{n})| \leq \gamma \qquad \forall \underline{n} \in \mathfrak{N}_0^m. \tag{2}$$

Note that the representation in (1) corresponds to the following shift operator polynomial:

$$P(q_1, \ldots, q_m) = 1 - \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}(\underline{n}) q_1^{i_1} \cdots q_m^{i_m}, \tag{3}$$

where q_i is the shift operator in the n_i direction. To guarantee the asymptotic stability of the system described by (1), it is necessary to show that the output approaches zero asymptotically along any direction for finite initial conditions. We adapt the definition of asymptotic stability given in [8] for use in the multidimensional case.

DEFINITION. The shift-variant m-D system in (1) is *asymptotically stable* in the region $\mathfrak{A}$ if and only if for any finite initial condition and any sequence of $\{\underline{a}(\underline{n})\} \in \mathfrak{A}$, the response $y(\underline{n})$ tends to zero asymptotically on $\mathfrak{IC}_K$ as $K \to \infty$.

Note that since we are dealing with linear systems, the concept of global asymptotic stability is equivalent to that of asymptotic stability. Note also that this definition places no restrictions on the rate of change of the coefficients. It also implies that if we can find one sequence of coefficient values contained in a region for which a nonconvergent response is obtained, that region is not a stable parameter region.

3. Main results

THEOREM 1. The linear shift-variant m-D system described by (1) and (2) is asymptotically stable if and only if $\gamma < 1$.

Proof. (*Sufficiency*) If we can show that the zero-input response of a m-D system is bounded along hyperplanes $\mathfrak{IC}_K$ and that the absolute bound on the response on each hyperplane converges to zero as K approaches infinity, we can conclude that the system is asymptotically stable.

Step 1. Note that for $\underline{n} \in \tilde{\mathcal{H}}_K$, one can always find a B sufficiently large such that $|y(\underline{n})| \leq B$. Furthermore, by choosing K sufficiently large, the initial conditions needed to compute output values on $\mathcal{H}_K$ are all zero. Now examine the bound on the output values for $\underline{n} \in \mathcal{H}_{K+1}$, given the condition stated in Theorem 1.

$$
\begin{aligned}
|y(\underline{n})|_{\underline{n} \in \mathcal{H}_{K+1}} &= \left| \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}) \right| \\
&\leq \sum_{\underline{i} \in \mathcal{I}} |a_{\underline{i}}(\underline{n})| \cdot |y(\underline{n} - \underline{i})| \\
&\leq B \cdot \sum_{\underline{i} \in \mathcal{I}} |a_{\underline{i}}(\underline{n})| \\
&\leq B \cdot \gamma.
\end{aligned}
$$

Since $\gamma < 1$, and $|y(\underline{n})| \leq B$ for all $\underline{n}$ satisfying $\underline{n} \in \tilde{\mathcal{H}}_K$, all outputs on any hyperplane $\mathcal{H}_{\tilde{K}+1}$ that satisfies $\tilde{K} \geq K$ are bounded in magnitude by $B \cdot \gamma$.

Step 2. Next we obtain a bound for $|y(\underline{n})|$ on the hyperplane $\mathcal{H}_{\tilde{K}+1}$ when $\tilde{K} \geq K + N_1 + \cdots + N_m$:

$$
\begin{aligned}
|y(\underline{n})|_{\underline{n} \in \mathcal{H}_{\tilde{K}+1}} &= \left| \sum_{\underline{i} \in \mathcal{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}) \right| \\
&\leq \sum_{\underline{i} \in \mathcal{I}} |a_{\underline{i}}(\underline{n})| \cdot |y(\underline{n} - \underline{i})| \\
&\leq B \cdot \gamma \cdot \sum_{\underline{i} \in \mathcal{I}} |a_{\underline{i}}(\underline{n})| \\
&\leq B \cdot \gamma^2.
\end{aligned}
$$

Step 3. By induction,

$$|y(\underline{n})| \leq B \cdot \gamma^{M+1}$$

for $\underline{n} \in \mathcal{H}_{\tilde{K}+1}$, where $\tilde{K} \geq K + M(N_1 + \cdots + N_m)$ and M is a nonnegative integer. Since from (2), we know that $\gamma < 1$, then as $M \to \infty$, $\gamma^M \to 0$, and the bound on $|y(\underline{n})|$ approaches 0.

(*Necessity*) If there exists a first hyperquadrant causal sequence of coefficients, $\{\underline{a}(\underline{n})\}$, $\underline{n} \in \mathfrak{N}_0^m$, which does not produce an asymptotically convergent system response $y(\underline{n})$, then the uncertain shift-variant system in (1) is unstable.

We will now show that if $\gamma \geq 1$ we can always produce a nonconvergent system response. Without loss of generality, consider the response $y(n_1, 0, \ldots, 0)$ along the n_1 axis. Choosing $a_i(\underline{n}) = 0$ for $i_j \neq 0$, $j = 2, \ldots, m$, we are left with an instantaneous output mask that is one-dimensional; i.e., the outputs propagate along the n_1 axis. We describe this system as follows:

$$y(n_1, 0, \ldots, 0) = \sum_{i_1=1}^{N_1} a_{i_1,0,\ldots,0}(n_1, 0, \ldots, 0)y(n_1 - i_1, 0, \ldots, 0).$$

Now if these coefficients satisfy

$$\sum_{i_1=1}^{N_1} |a_{i_1,0,\ldots 0}(n_1, 0, \ldots, 0)| \leq \gamma, \qquad \gamma \geq 1,$$

a system which is not asymptotically stable can be constructed by the 1-D result given in [8]. This is true because, since $\gamma \geq 1$, it is possible to select a sequence of coefficient values that prevents this response along the n_1 axis from converging to zero. (Asymptotic stability requires the response to converge to zero in all possible directions.)

Theorem 1 describes a necessary and sufficient condition for the asymptotic stability of (1) and (2) that holds for all $\underline{n} \in \mathfrak{N}_0^m$. The region $\mathfrak{A}$ defined by this condition is a closed high-dimensional hyperdiamond centered at the origin in the interior of the unit hyperdiamond.

Furthermore, $\gamma < 1$ limits the size of the largest stable hyperdiamond in the coefficient space that is fully symmetric around zero. This is true for shift-variant and shift-invariant systems.

Note that while Theorem 1 addresses the stability of shift-variant m-D systems, we can also specify the largest stable hyperdiamond for the *shift-invariant* case:

$$\sum_{\underline{i} \in \mathfrak{I}} |a_{\underline{i}}| \leq \gamma < 1. \tag{4}$$

To show that the condition on γ is necessary, consider the system

$$y(\underline{n}) = \sum_{\underline{i} \in \mathfrak{I}} a_{\underline{i}} y(\underline{n} - \underline{i}).$$

If the coefficients can be chosen from the hyperdiamond described in (4), then we can certainly choose $a_{1,0,\ldots,0} = \gamma$ and $a_{\underline{i}} = 0$ for all other $\underline{i} \in \mathfrak{I}$. Clearly, γ must be less than 1 for the system response to converge to zero.

By using information about the previous system outputs, a less restrictive shift-variant sequence of regions $\mathfrak{A}(\underline{n})$ in the parameter space can be identified for each $\underline{n} \in \mathfrak{N}_0^m$.

THEOREM 2. The m-D system described by (1) is asymptotically stable if

$$\mathfrak{A}(\underline{n}) = \left\{ \underline{a}(\underline{n}) \in \mathfrak{A}(\underline{n}) : -\gamma \cdot \max_{\underline{i} \in \mathfrak{I}} \{|y(\underline{n} - \underline{i})|\} \right.$$

$$\left. \leq \sum_{\underline{i} \in \mathfrak{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}) \leq \gamma \cdot \max_{\underline{i} \in \mathfrak{I}} \{|y(\underline{n} - \underline{i})|\} \right\}, \tag{5}$$

where $0 \leq \gamma < 1$.

Proof. If $\underline{a}(\underline{n})$ is chosen such that

$$|y(\underline{n})| \leq \gamma \cdot \max_{\underline{i} \in \mathfrak{I}} \{|y(\underline{n} - \underline{i})|\}, \qquad 0 < \gamma < 1, \tag{6}$$

we can construct an exponentially decaying upper bound on the response $y(\underline{n})$. In fact, using (1) to make a substitution for $y(\underline{n})$ in (6), we have

$$\left| \sum_{\underline{i} \in \mathfrak{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}) \right| \leq \gamma \cdot \max_{\underline{i} \in \mathfrak{I}} \{|y(\underline{n} - \underline{i})|\}, \qquad 0 < \gamma < 1. \tag{7}$$

This means that we can express $\mathfrak{A}(n)$, the shift-variant region of stability, as the intersection of two half-spaces with parallel boundaries:

$$-\gamma \cdot \max_{\underline{i} \in \mathfrak{I}} \{|y(\underline{n} - \underline{i})|\} \leq \sum_{\underline{i} \in \mathfrak{I}} a_{\underline{i}}(\underline{n}) y(\underline{n} - \underline{i}) \leq \gamma \cdot \max_{\underline{i} \in \mathfrak{I}} \{|y(\underline{n} - \underline{i})|\}. \tag{8}$$

Remarks. The region $\mathfrak{A}(\underline{n})$ consists of the intersection of two half-spaces with parallel boundaries in the coefficient space, and the volume of this region is infinite. One could think of $\underline{a}(\underline{n})$ as a conveniently ordered vector of the m-D shift-variant coefficients in (1) and $\underline{y}(\underline{n})$ as the corresponding vector of previous outputs. We can then express (1) as an inner product of $\underline{a}(\underline{n})$ and $\underline{y}(\underline{n})$. So we can speak of the boundaries of the intersecting half-spaces in the coefficient space as being orthogonal to $\underline{y}(\underline{n})$. Note that the construction of the shift-variant region of Theorem 2 requires information about previous system outputs. This region always includes the shift-invariant region of Theorem 1. In the case where we have bounded coefficients, input-output stability is also guaranteed because we have exponential zero convergence of the zero-input response [17].

This result has important consequences in the field of adaptive filtering. The construction of the shift-variant region allows for a reduction in error in the choice of updated coefficients in an adaptive system. Usually, the computed coefficients are projected inside the stable region of the shift-invariant system [18]. Using the shift-variant region of Theorem 2, the error due to projection could be significantly reduced compared to the case which requires projection into the parameter region of shift-invariant system stability.

4. Conclusion

This paper addressed the problem of asymptotic stability of shift-variant multidimensional difference equations. The largest shift-invariant region of stability that is symmetric with respect to zero in the coefficient space was found. A shift-variant sequence of regions of stability with infinite volume was also constructed. These results extend the 1-D results of [8]. For adaptive filtering applications, the results allow reduced error in choosing updated filter coefficients.

The work here also introduced a useful methodology for analyzing the response behavior of m-D systems. The usefulness of the concept of absolute response bounds is not limited to stability analysis. The techniques used throughout this paper may also be useful in assessing system performance.

Acknowledgment

This work was supported in part by funds from the Clare Boothe Luce Foundation, the Society of Automotive Engineers, and the Office of Naval Research Grant #N00014-94-1-0387.

References

1. E.I. Jury, "Stability of Multidimensional Systems and Related Problems," in *Multidimensional Systems: Techniques and Applications* (S.G. Tzafestas, ed.), New York: Marcel Dekker, 1986, Chap. 3.
2. C.V. Hollot and A.C. Bartlett, "Some Discrete-Time Counterparts to Kharitonov's Stability Criterion for Uncertain Systems," *IEEE Trans. Automat. Control*, vol. 31, 1986, pp. 355–356.
3. F. Kraus, B.D.O. Anderson, E.I. Jury, and M. Mansour, "On the Robustness of Low-Order Schur Polynomials," *IEEE Trans. Circuits Systems*, 1988, vol. 35, pp. 570–577.
4. J.E. Ackerman and B.R. Barmish, "Robust Schur Stability of a Polytope of Polynomials," *IEEE Trans. Automat. Control*, vol. 33, 1988, pp. 984–986.
5. F.J. Kraus and M. Mansour, "On Robust Stability of Discrete Systems," in *Proc. 29th IEEE Conf. Decision Contr.*, Honolulu, HI, 1990, pp. 421–422.
6. Y.K. Foo and Y.C. Soh, "Schur Stability of Interval Polynomials," *IEEE Trans. Automat. Control*, vol. 38, 1993, pp. 943–946.
7. V.L. Kharitonov, "Asymptotic Stability of an Equilibrium Position of a Family of Systems of Linear Differential Equations," *Differential'nye Uravnenia*, vol. 14, 1978, pp. 2086–2088.
8. P.H. Bauer, M. Mansour, and J. Durán, "Stability of Polynomials with Time-Variant Coefficients," *IEEE Trans. Circuits Systems*, vol. 40, 1993, pp. 423–426.
9. P.H. Bauer, K. Premaratne, and J. Durán, "A Necessary and Sufficient Condition for Robust Asymptotic Stability of Time-Variant Discrete Systems," *IEEE Trans. Automat. Control*, vol. 38, 1993, pp. 1427–1430.
10. E. Yaz and X. Niu, "Stability Robustness of Linear Discrete-Time Systems in the Presence of Uncertainties," *IJC*, vol. 50, 1989, pp. 173–182.
11. S.R. Kolla, R.A. Yedavalli, and J.B. Farison, "Robust Stability Bounds of Time-Varying Perturbations for State Space Models of Linear Discrete-Time Systems," *IJC*, vol. 50, 1989, pp. 151–159.
12. W.-S. Lu, "2-D Stability Test via 1-D Stability Robustness Analysis," *IJC*, vol. 48, 1988, pp. 1735–1741.
13. P.H. Bauer and K. Premaratne, "Robust Stability of Time-Variant Interval Matrices," in *Proc. 29th IEEE Conf. Decision Contr.*, Honolulu, HI, 1990, pp. 434–435.
14. P.H. Bauer and E.I. Jury, "BIBO-Stability of Multidimensional (mD) Shift-Variant Discrete Systems," *IEEE Trans. Automat. Control*, vol. 36, 1991, pp. 1057–1061.

15. P.H. Bauer and E.I. Jury, "A Stability Analysis of Two-Dimensional Nonlinear Digital State-Space Filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, 1990, pp. 1578–1586.
16. P.H. Bauer, "Robustness and Stability Properties of First-Order Multidimensional (m-D) Discrete Processes," *Multidimens. Syst. Signal Proc.*, 1990, vol. 1, pp. 75–86.
17. L.M. Silverman and B.D.O. Anderson, "Controllability, Observability and Stability of Linear Systems," *SIAM J. Control*, vol. 6, 1968, pp. 121–130.
18. C.R. Johnson, "Adaptive IIR Filtering: Current Results and Open Issues," *IEEE Trans. Inform. Theory*, vol. IT-30, 1984, pp. 237–250.

OFFICE OF THE UNDER SECRETARY OF DEFENSE (ACQUISITION)
DEFENSE TECHNICAL INFORMATION CENTER
CAMERON STATION
ALEXANDRIA, VIRGINIA 22304-6145

IN REPLY REFER TO DTIC-OCC

SUBJECT: Distribution Statements on Technical Documents

TO: OFFICE OF NAVAL RESEARCH
CORPORATE PROGRAMS DIVISION
ONR 353
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5660

1995 1031 044

1. Reference: DoD Directive 5230.24, Distribution Statements on Technical Documents, 18 Mar 87.

2. The Defense Technical Information Center received the enclosed report (referenced below) which is not marked in accordance with the above reference.

FINAL REPORT
N00014-94-1-0387
TITLE: HIGH-SPEED FIXED-AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE TIME SYSTEMS

3. We request the appropriate distribution statement be assigned and the report returned to DTIC within 5 working days.

4. Approved distribution statements are listed on the reverse of this letter. If you have any questions regarding these statements, call DTIC's Cataloging Branch, (703) 274-6837.

FOR THE ADMINISTRATOR:

1 Encl

GOPALAKRISHNAN NAIR
Chief, Cataloging Branch

FL-171
Jul 93

DISTRIBUTION STATEMENT A:

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

DISTRIBUTION STATEMENT B:

DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES ONLY; (Indicate Reason and Date Below). OTHER REQUESTS FOR THIS DOCUMENT SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT C:

DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES AND THEIR CONTRACTORS; (Indicate Reason and Date Below). OTHER REQUESTS FOR THIS DOCUMENT SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT D:

DISTRIBUTION AUTHORIZED TO DOD AND U.S. DOD CONTRACTORS ONLY; (Indicate Reason and Date Below). OTHER REQUESTS SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT E:

DISTRIBUTION AUTHORIZED TO DOD COMPONENTS ONLY; (Indicate Reason and Date Below). OTHER REQUESTS SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT F:

FURTHER DISSEMINATION ONLY AS DIRECTED BY (Indicate Controlling DoD Office and Date Below) or HIGHER DOD AUTHORITY.

DISTRIBUTION STATEMENT X:

DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES AND PRIVATE INDIVIDUALS OR ENTERPRISES ELIGIBLE TO OBTAIN EXPORT-CONTROLLED TECHNICAL DATA IN ACCORDANCE WITH DOD DIRECTIVE 5230.25, WITHHOLDING OF UNCLASSIFIED TECHNICAL DATA FROM PUBLIC DISCLOSURE, 6 Nov 1984 (Indicate date of determination). CONTROLLING DOD OFFICE IS (Indicate Controlling DoD Office).

The cited documents has been reviewed by competent authority and the following distribution statement is hereby authorized.

A

(Statement)

OFFICE OF NAVAL RESEARCH
CORPORATE PROGRAMS DIVISION
ONR 353
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5660

(Controlling DoD Office Name)

(Reason)

(Controlling DoD Office Address, City, State, Zip)

DEBRA T. HUGHES
DEPUTY DIRECTOR
CORPORATE PROGRAMS OFFICE

(Signature & Typed Name)

(Assigning Office)

19 SEP 1995

(Date Statement Assigned)